SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

ATTRACTORS FOR THE PENALIZED NAVIER–STOKES EQUATIONS*

B. BREFORT†, J. M. GHIDAGLIA† AND R. TEMAM†

Abstract. We consider the penalized form of the Navier–Stokes equations for a viscous incompressible fluid where the pressure and the incompressibility equation div $u = 0$ are suppressed and replaced by a penalty term in the momentum conservation equation. In this article we study the existence of an attractor for the penalized Navier–Stokes equation, this attractor describing the long-time behaviour of the solutions. Then we let the penalty parameter tend to zero and we show how the attractors of the penalized equations approximate the attractor of the exact equations.

Key words. attractors, Navier–Stokes equations, penalization

AMS(MOS) subject classifications. 35Q10, 65P05, 76D05

Introduction. The penalty method was introduced by Courant [5] in the context of the calculus of variations and has developed considerably. Besides the applications to the constrained variational problems and variational inequalities, this is now a useful tool for numerical computations in continuum fluid and solid mechanics. In particular its application to the Navier–Stokes equations which, as far as we know, was initiated in Temam [13], [14] is now commonly used in some areas of computational fluid dynamics, especially for the computations using the finite element methods in conjunction with quadrature formula (see for instance Bercovier [2], Bercovier and Engelman [3], Oden and Kikuchi [11], Oden and Jacquotte [12]).

With the increase of the computing power we are now at the point of being able to compute nonstationary flows; by this we mean the time periodic flows or the more complex (turbulent) flows which do appear, even if the driving forces are time-independent, after a Hopf bifurcation or a cascade of more complex bifurcations has occurred (Feigenbaum cascade of bifurcations, Ruelle–Takens bifurcations towards turbulence, ...). This new development in computational fluid dynamics will necessitate some improvement of our knowledge of the dynamics of the Navier–Stokes equations, and the problem that we address here pertains to this question.

Our aim in this article is to study the attractors for the penalized Navier–Stokes Equations (N.S.E.) and their convergence toward the universal attractor of the exact equations (cf. Foias and Temam [6], Temam [18], and the references therein). We restrict ourselves to the two-dimensional case and consider the flow in a bounded domain $\Omega \subset \mathbb{R}^2$; the N.S.E. of incompressible flows then reads

$$\frac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla) u + \operatorname{grad} p = f, \tag{0.1}$$

$$\operatorname{div} u = 0 \quad \text{in } \Omega \times (0, T) \tag{0.2}$$

where $u = u(x, t)$ is the velocity vector, $p = p(x, t)$ the pressure, $\nu > 0$ is the kinematic viscosity and f represents the volumic driving forces, for simplicity the constant density ρ was taken equal to 1.

For the penalized equation we suppress the pressure p and the incompressibility equation (0.2) and introduce in (0.1) a penalty term, (ν/ε) grad div u, $\varepsilon > 0$ the penalty

* Received by the editors September 3, 1986; accepted for publication October 29, 1986.

† Laboratoire d'Analyse Numérique, Centre National de la Recherche Scientifique, Université Paris-Sud, Bât. 425, 91405-Orsay, France.

parameters. Hence we obtain[1]

$$\frac{\partial u_\varepsilon}{\partial t} - \nu\Delta u_\varepsilon + (u_\varepsilon \cdot \nabla)u_\varepsilon + \frac{1}{2}(\operatorname{div} u_\varepsilon)u_\varepsilon - \frac{\nu}{\varepsilon}\operatorname{grad}\operatorname{div} u_\varepsilon = f. \tag{0.3}$$

We have also introduced the supplementary nonlinear term $\frac{1}{2}(\operatorname{div} u_\varepsilon)u_\varepsilon$ which was proposed in [13], [14] to make (0.3) well set.

The equations (0.1), (0.2) (or (0.3)) are supplemented by boundary and initial conditions. For the boundary condition, two cases will be considered:

Either Ω is a smooth bounded domain of $\mathbb{R}^2$ with boundary Γ and we set

$$u = 0 \quad \text{on } \Gamma \times (0, T), \tag{0.4}$$

Or Ω is a square in $\mathbb{R}^2$, $\Omega = (0, L)^2$, and the boundary condition is the space periodicity for u and p

$$\begin{aligned} &\varphi(x_1, L, t) = \varphi(x_1, 0, t), \\ &\varphi(L, x_2, t) = \varphi(0, x_2, t), \\ &0 < x_1, x_2 < L, \quad t \in (0, T), \quad \varphi = u \text{ or } p. \end{aligned} \tag{0.5}$$

In this case we assume also that the average flow in Ω vanishes (cf. [6])

$$\int_\Omega u(x, t)\, dx = 0 \quad \forall t. \tag{0.6}$$

The initial condition is simply in all cases

$$u(x, 0) = u_0(x), \qquad x \in \Omega. \tag{0.7}$$

This article is organized as follows. In § 1 we recall the functional form of the exact and penalized N.S.E. and the results of existence, uniqueness and regularity of solutions in both cases (§ 1.1); § 1.2 contains a technical result on the penalized Stokes problem. Section 2 is devoted to the study of the attractors of the penalized N.S.E. (the penalization parameter ε being fixed). We derive uniform estimates for various norms of u and prove the existence of an absorbing set and a universal attractor $\mathscr{A}_\varepsilon$ for all $\varepsilon > 0$. Then § 3 deals with the convergence of the attractor $\mathscr{A}_\varepsilon$ to the universal attractor $\mathscr{A}$ of the exact Navier-Stokes equations when $\varepsilon \to 0$. Section 3.1 provides an appropriate result on convergence of attractors; § 3.2 recalls and improves the results of [13], [14] of convergence of the solutions u_ε of the penalized N.S.E. to the solution u of the exact N.S.E. Section 3.3 gives the main result of convergence of $\mathscr{A}_\varepsilon$ to $\mathscr{A}$. Finally § 3.4 shows how the dimension of $\mathscr{A}_\varepsilon$ can be compared to that of $\mathscr{A}$.

1. Survey and complements for the exact and penalized Navier–Stokes equations.

1.1. Functional setting. We denote by $L^2(\Omega)$ the space of square integrable functions on Ω, and by $H^m(\Omega)$ the Sobolev space of order m based on $L^2(\Omega)$ (m an integer), i.e., $H^m(\Omega)$ is the space of functions u in $L^2(\Omega)$ whose distribution derivatives of order $\leqq m$ are in $L^2(\Omega)$. We let $H_0^1(\Omega)$ denote the space of functions in $H^1(\Omega)$ whose trace on Γ vanishes and in the periodic case (0.5) $H^1_{\text{per}}(\Omega)$ represents the space of functions u in $H^1(\Omega)$ whose trace assumes the same values on corresponding points of Γ. Finally we set $\mathbb{L}^2(\Omega) = L^2(\Omega)^n$, $\mathbb{H}^1(\Omega) = H^1(\Omega)^n$, etc. $\cdots$, $n = 2$. The scalar products and norms on either $L^2(\Omega)$ or $\mathbb{L}^2(\Omega)$ are denoted

$$(u, v) = \int_\Omega u(x) \cdot v(x)\, dx, \qquad |u| = \{(u, u)\}^{1/2} \tag{1.1}$$

[1] The penalty parameter is written ε/ν instead of ε as in [14]. Hence ε is a nondimensional parameter.

and on $H_0^1(\Omega)$ or $\mathbb{H}_0^1(\Omega)$:

$$((u, v)) = \sum_{i=1}^{n} (D_i u, D_i v), \quad \|u\| = \{((u, u))\}^{1/2} \quad \left(D_i = \frac{\partial}{\partial x_i}\right). \tag{1.2}$$

We have the Poincaré inequality

$$|u| \leqq \frac{1}{\sqrt{\lambda_1}} \|u\| \quad \forall u \in H_0^1(\Omega) \text{ (or } \mathbb{H}_0^1(\Omega)), \tag{1.3}$$

where $\lambda_1 > 0$ is the first eigenvalue of the Laplace operator in $H_0^1(\Omega)$; (1.3) shows that $\|u\|$ is a norm on $H_0^1(\Omega)$ (or $\mathbb{H}_0^1(\Omega)$) which is equivalent to that of $H^1(\Omega)$ (or $\mathbb{H}^1(\Omega)$).

A similar remark holds in the space periodic case if we restrict ourselves to functions satisfying (0.6):

$$\dot{L}^2(\Omega) = \{u \in L^2(\Omega), u \text{ satisfies (0.6)}\},$$

$$\dot{H}^1(\Omega) = \{u \in H^1(\Omega), u \text{ satisfies (0.6)}\}, \cdots.$$

In particular we have

$$|u| \leqq \frac{1}{\sqrt{\lambda_1'}} \|u\| \quad \forall u \in H^1_{\text{per}}(\Omega) \text{ (or } \dot{\mathbb{H}}^1_{\text{per}}(\Omega)) \tag{1.4}$$

where $\lambda_1' = 4\pi^2/L^2$ is the first eigenvalue of the Laplace operator in $\dot{H}^1_{\text{per}}(\Omega)$ (or $\mathbb{H}^1_{\text{per}}(\Omega)$). This shows that in this case too $\|u\|$ is a norm on $\dot{H}^1_{\text{per}}(\Omega)$ (or $\mathbb{H}^1_{\text{per}}(\Omega)$) equivalent to that of $H^1(\Omega)$ (or $\mathbb{H}^1(\Omega)$).

The basic Hilbert space is $H \subset \mathbb{L}^2(\Omega)$; in the case (0.4) (cf. [15]),

$$H = \{u \in \mathbb{L}^2(\Omega), \operatorname{div} u = 0, u \cdot \underset{\sim}{\nu} = 0 \text{ on } \Gamma\}$$

where $\underset{\sim}{\nu}$ is the unit outward normal on Γ and in the case (0.5) (cf. [16]), H is the set of $u \in \mathbb{L}^2(\Omega)$ such that $\operatorname{div} u = 0$ and $u \cdot \underset{\sim}{\nu}$ take opposite values on corresponding points of Γ. We denote by A the unbounded operator in H with domain

$$D(A) = \{u \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega), \operatorname{div} u = 0\} \quad \text{in the case (0.4)}, \tag{1.5}$$

$$D(A) = \{u \in \dot{\mathbb{H}}^2_{\text{per}}(\Omega), \operatorname{div} u = 0\} \quad \text{in the case (0.5)}, \tag{1.6}$$

and

$$Au = -P\Delta u \quad \forall u \in D(A),$$

where P is the projector in $\mathbb{L}^2(\Omega)$ onto H. The operator A is self-adjoint, >0 in H, and A is an isomorphism from $D(A)$ (endowed with the graph norm) onto H. Since by Rellich's theorem the embedding of $H^1(\Omega)$ into $L^2(\Omega)$ is compact, the embedding of $D(A)$ in H is compact and therefore A^{-1} is a compact self-adjoint operator in H. Thus there exists an orthonormal basis of H consisting of eigenvectors w_j of A (or A^{-1}),

$$\begin{cases} Aw_j = \lambda_j w_j \quad \text{(for (0.4))}, \\ Aw_j' = \lambda_j' w_j' \quad \text{(for (0.5))}, \\ 0 < \lambda_1 \leqq \lambda_2, \cdots, \qquad 0 < \lambda_1' \leqq \lambda_2', \cdots, \\ \lambda_j, \lambda_j' \to \infty \quad \text{as } j \to \infty. \end{cases} \tag{1.7}$$

We can define the powers A^α of A, $\alpha \in \mathbb{R}$. Of particular interest are the spaces $V = D(A^{1/2})$, $V' = D(A^{-1/2})$ = the dual of V, and [15], [16],

$$V = \{u \in \mathbb{H}_0^1(\Omega), \operatorname{div} u = 0\} \quad \text{(for (0.4))}, \tag{1.8}$$

$$V = \{u \in \mathbb{H}_{\text{per}}^1(\Omega), \operatorname{div} u = 0\} \quad \text{(for (0.5))}. \tag{1.9}$$

The space V is Hilbert for the scalar product $((u, v))$.

We are given $f \in \mathbb{L}^2(\Omega)^2$ and for $u, v \in V$ or $\mathbb{H}^1(\Omega)$, we denote by $B(u, v)$ the element of V' defined by

$$(B(u, v), w) = \sum_{i,j=1}^{2} \int_\Omega u_i (D_i v_j) w_j \, dx \quad \forall w \in V \tag{1.10}$$

and we set $B(u) = B(u, u)$.

Similarly for the penalized problems the basic spaces are $G = \mathbb{L}^2(\Omega)$ in the case (0.4), $= \dot{\mathbb{L}}^2(\Omega)$ in the case (0.5), and $W = \mathbb{H}_0^1(\Omega)$ (for (0.4)) or $\dot{\mathbb{H}}_{\text{per}}^1(\Omega)$ (for (0.5)). The operator A is replaced by $-\Delta$, $D(-\Delta) = \mathbb{H}_{\text{per}}^2(\Omega) \cap W$ and for $u, v \in \mathbb{H}^1(\Omega)$, we denote by $\mathcal{B}(u, v)$ and $\tilde{\mathcal{B}}(u, v)$ the elements of W' defined by

$$(\mathcal{B}(u, v), w) = \sum_{i,j=1}^{2} \int_\Omega u_i (D_i v_j) w_j \, dx \quad \forall w \in W, \tag{1.11}$$

$$\begin{aligned} (\tilde{\mathcal{B}}(u, v), w) &= (\mathcal{B}(u, v), w) + \frac{1}{2} \int_\Omega (\operatorname{div} u) v w \, dx \\ &= \text{(cf. [14], [15])} \\ &= \frac{1}{2} \sum_{i,j=1}^{2} \int_\Omega u_i \{(D_i v_j) w_j - v_j (D_i w_j)\} \, dx \quad \forall w \in W. \end{aligned} \tag{1.12}$$

We also set $\mathcal{B}(u) = \mathcal{B}(u, u)$, $\tilde{\mathcal{B}}(u) = \tilde{\mathcal{B}}(u, u)$.

The functional form of the exact N.S.E. is then (cf. [15], [16])

$$\frac{du}{dt} + \nu A u + B(u) = Pf \quad \text{in } V' \tag{1.13}$$

which we supplement with the initial condition (for initial value problems).

$$u(0) = u_0. \tag{1.14}$$

The penalized N.S.E. are written

$$\frac{du_\varepsilon}{dt} - \nu \Delta u_\varepsilon + \tilde{\mathcal{B}}(u_\varepsilon) - \frac{\nu}{\varepsilon} \operatorname{grad} \operatorname{div} u_\varepsilon = f \quad \text{in } W' \tag{1.15}$$

with an initial condition

$$u_\varepsilon(0) = u_{\varepsilon 0}. \tag{1.16}$$

We recall that for u_0 given in H (and $Pf \in H$), (1.13), (1.14) possesses a unique solution $u \in L^2(0, T, V) \cap \mathcal{C}([0, T]; H)$. In fact, [6], u is analytic from $]0, \infty[$ with values in $D(A)$ and if $u_0 \in V$ (instead of H), $u \in L^2_{\text{loc}}(0, \infty; D(A)) \cap \mathcal{C}([0, \infty[; V)$.

[2] We restrict ourselves to f independent of time since we consider only autonomous dynamical systems.

For (1.15), (1.16) similar results were proved in [13], [14]; there exists a unique solution $u_\varepsilon \in L^2(0, T, W) \cap L^\infty([0, T]; G)$ for all $T > 0$, of (1.15), (1.16). Also with the same methods as in [6], we find that u_ε is analytic in t from $(0, \infty)$ with values in $D(-\Delta)$ = the domain of $-\Delta$ in G (= a closed subspace of $\mathbb{H}^2(\Omega)$). Furthermore if $u_{\varepsilon 0} \in W$, then u_ε is in $L^2_{\text{loc}}(0, \infty; \mathbb{H}^2(\Omega)) \cap \mathscr{C}([0, \infty[; W)$. We denote by $S(t)$ the operator $u_0 \to u(t)$ which maps H into $D(A)$ for $t > 0$, and by $S_\varepsilon(t)$ the operator $u_{\varepsilon 0} \to u_\varepsilon(t)$ which maps $\mathbb{L}^2(\Omega)$ into $\mathbb{H}^2(\Omega)$. Each of the family of operators $\{S(t)\}_{t>0}$, $\{S_\varepsilon(t)\}_{t \geqq 0}$ enjoys the usual semigroup properties.

The results above are valid for each $\varepsilon > 0$ fixed. Concerning the passage to the limit $\varepsilon \to 0$ some convergence results of u_ε to u were proved in [13], [14]. We will recall and complete these results in § 3. We now finish this section with a technical result.

1.2. Comparison of two norms. We consider mainly in this section the case of Dirichlet boundary condition (0.4), while the space periodic case (0.5) is much easier and will be rapidly considered at the end of this section.

We recall [15] that, for $u \in D(A)$ (or even V) and $g \in H$, saying that $Av = g$ amounts to saying that there exists $q \in L^2(\Omega)$ such that

$$\begin{aligned} -\Delta v + \operatorname{grad} q &= g \quad \text{in } \Omega, \\ \operatorname{div} v &= 0 \quad \text{in } \Omega, \\ v &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{1.17}$$

This is a Stokes problem. We can also consider a slightly more general Stokes problem

$$\begin{aligned} -\Delta v + \operatorname{grad} q &= g \quad \text{in } \Omega, \\ \operatorname{div} v &= h \quad \text{in } \Omega, \\ v &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{1.18}$$

Now $g \in H$ or more generally $\mathbb{L}^2(\Omega)$, and for instance $h \in \dot{L}^2(\Omega)$. The results concerning (1.18) can be found in [15] and will be recalled when needed.

The penalized form of (1.17) is the following [15]; given $\varepsilon > 0$ and $g_\varepsilon \in \mathbb{L}^2(\Omega)$, we denote by v_ε the solution of

$$\begin{aligned} -\Delta v_\varepsilon - \frac{1}{\varepsilon} \operatorname{grad} \operatorname{div} v_\varepsilon &= g_\varepsilon \quad \text{in } \Omega, \\ v_\varepsilon &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{1.19}$$

It is known that if $v \in V$ and $Av = g \in H$ then in fact $v \in D(A) = \mathbb{H}^2(\Omega) \cap V$ and $|Av|$ is a norm on $D(A)$ which is equivalent to that of $\mathbb{H}^2(\Omega)$. More generally for (1.18), if g is in $\mathbb{L}^2(\Omega)$ and h in $H^1(\Omega)$ then v is in $\mathbb{H}^2(\Omega)$, q in $H^1(\Omega)$ and there exists a constant C_1 depending only on Ω such that[3]

$$\|v\|_{\mathbb{H}^2(\Omega)} + |q|_{H^1(\Omega)} \leqq C_1(|g|_{\mathbb{L}^2(\Omega)} + |h|_{H^1(\Omega)}) \tag{1.20}$$

or equivalently with another constant C_2

$$|\Delta v|_{\mathbb{L}^2(\Omega)} + |\operatorname{grad} q|_{\mathbb{L}^2(\Omega)} \leqq C_2(|g|_{\mathbb{L}^2(\Omega)} + |\operatorname{grad} h|_{\mathbb{L}^2(\Omega)}). \tag{1.21}$$

For (1.19) the similar results are simple and follow immediately from the general results on elliptic systems (cf. Agmon, Douglis and Nirenberg [1]). For fixed $\varepsilon > 0$, if

[3] Here and in the sequel we make q unique by imposing the condition $\int_\Omega q(x)\, dx = 0$.

$g_\varepsilon \in \mathbb{L}^2(\Omega)$ and $v_\varepsilon \in \mathbb{H}_0^1(\Omega)$ satisfies (1.19) then in fact $v_\varepsilon \in \mathbb{H}^2(\Omega)$ and $|\mathcal{A}_\varepsilon v|$ is a norm on $\mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega)$, equivalent to that of $\mathbb{H}^2(\Omega)$; here we have written

$$\mathcal{A}_\varepsilon v = -\Delta v - \frac{1}{\varepsilon} \text{grad div } v \quad \forall v \in \mathbb{H}_0^1(\Omega).$$

Of course [1] provides only an equivalence of norms

$$|\mathcal{A}_\varepsilon v| \sim |v|_{\mathbb{H}^2(\Omega)} \tag{1.22}$$

with constants depending on ε (besides Ω). Our aim is now to show that one of the inequalities (1.22) is valid uniformly with respect to ε.

LEMMA 1.1. *There exists a constant $C_3 < \infty$, depending only on Ω and such that if $\varepsilon C_3 \leqq 1$*

$$|\Delta v| \leqq C_3 |\mathcal{A}_\varepsilon v| \quad \forall v \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega). \tag{1.23}$$

Proof. Let $v \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega)$ and let $g_\varepsilon = \mathcal{A}_\varepsilon v$. Taking the scalar product with v we obtain

$$\begin{aligned} \|v\|^2 + \frac{1}{\varepsilon} |\text{div } v|^2 &= (g_\varepsilon, v) \\ &\leqq |g_\varepsilon| |v| \leqq \frac{1}{\sqrt{\lambda_1}} |g_\varepsilon| \|v\| \\ &\leqq \frac{1}{2} \|v\|^2 + \frac{1}{2\lambda_1} |g_\varepsilon|^2. \end{aligned}$$

Hence

$$\|v\|^2 + \frac{1}{\varepsilon} |\text{div } v|^2 \leqq \frac{1}{\lambda_1} |g_\varepsilon|^2 = \frac{1}{\lambda_1} |\mathcal{A}_\varepsilon v|^2. \tag{1.24}$$

Now setting $q_\varepsilon = (1/\varepsilon) \text{ div } v$, we rewrite $\mathcal{A}_\varepsilon v = g_\varepsilon$ in the following manner

$$\begin{aligned} -\Delta v + \text{grad } q_\varepsilon &= g_\varepsilon, \\ \text{div } v &= h_\varepsilon. \end{aligned}$$

Using (1.21) we then find

$$|\Delta v| + |\text{grad } q_\varepsilon| \leqq C_2 (|g_\varepsilon| + |\text{grad } h_\varepsilon|)$$

or equivalently

$$|\Delta v| + \frac{1}{\varepsilon} |\text{grad div } v| \leqq C_2 (|g_\varepsilon| + |\text{grad div } v|)$$

and if

$$\varepsilon C_2 \leqq \tfrac{1}{2}, \tag{1.25}$$

$$|\Delta v| + \frac{1}{\varepsilon} |\text{grad div } v| \leqq 2C_2 |g_\varepsilon|. \tag{1.26}$$

This proves (1.23) with $C_3 = 2C_2$. We recall that $|\Delta v|$ and $|v|_{\mathbb{H}^2(\Omega)}$ are equivalent norms on $\mathbb{H}^2(\Omega) \cap \mathbb{H}_0^1(\Omega)$.

Remark 1.1. We infer also from (1.26) that

$$|\text{grad div } v| \leqq \varepsilon C_3 |\mathcal{A}_\varepsilon v|. \tag{1.27}$$

Remark 1.2. In the space periodic case, an inequality similar to (1.23) holds, for every $v \in \dot{\mathbb{H}}^2_{\text{per}}(\Omega)$. The proof is exactly the same as Lemma 1.1 and relies on the analogue of (1.21), which is much easier to prove in this case by using Fourier series expansions [16].

2. Attractors for the penalized equations. We derive in §§ 2.1 and 2.2 a series of uniform estimates on the solutions of the penalized N.S.E. We then introduce the absorbing sets and their universal attractor.

2.1. Uniform estimates in L^2. If u_ε is solution of (1.15), (1.16) then, taking the scalar product of (1.15) with u_ε in $\mathbb{L}^2(\Omega)$ we obtain

$$\frac{1}{2}\frac{d}{dt}|u_\varepsilon|^2 + \nu\|u_\varepsilon\|^2 + \frac{\nu}{\varepsilon}|\text{div } u_\varepsilon|^2 = (f, u_\varepsilon)$$

$$\leqq |f||u_\varepsilon| \leqq \frac{1}{\sqrt{\lambda_1}}|f|\,\|u_\varepsilon\|$$

$$\leqq \frac{\nu}{2}\|u_\varepsilon\|^2 + \frac{1}{2\nu\lambda_1}|f|^2.$$

We have used the property

$$(\tilde{\mathcal{B}}(u, v), v) = 0 \quad \forall u, v \in W \tag{2.1}$$

which follows easily from the second expression of $\tilde{\mathcal{B}}$ in (1.12). Therefore

$$\frac{d}{dt}|u_\varepsilon|^2 + \nu\|u_\varepsilon\|^2 + \frac{\nu}{\varepsilon}|\text{div } u_\varepsilon|^2 \leqq \frac{1}{2\lambda_1}|f|^2, \tag{2.2}$$

$$\frac{d}{dt}|u_\varepsilon|^2 + \nu\lambda_1|u_\varepsilon|^2 + \frac{\nu}{\varepsilon}|\text{div } u_\varepsilon|^2 \leqq \frac{1}{\nu\lambda_1}|f|^2, \tag{2.3}$$

from which we infer easily that

$$|u_\varepsilon(t)|^2 \leqq |u_{\varepsilon 0}|^2\, e^{-\nu\lambda_1 t} + \frac{1}{\nu^2\lambda_1^2}|f|^2(1 - e^{-\nu\lambda_1 t}), \tag{2.4}$$

$$\overline{\lim_{t\to\infty}}\ |u_\varepsilon(t)|^2 \leqq \frac{1}{\nu^2\lambda_1^2}|f|^2. \tag{2.5}$$

Returning then to (2.2) we obtain

$$\nu\int_t^{t+T}\left\{\nu\|u_\varepsilon\|^2 + \frac{1}{\varepsilon}|\text{div } u_\varepsilon|^2\right\} ds \leqq |u_{\varepsilon 0}|^2\, e^{-\nu\lambda_1 t} + \frac{T}{\nu\lambda_1}|f|^2 + \frac{|f|^2}{\nu^2\lambda_1^2}(1 - e^{-\nu\lambda_1 t}), \tag{2.6}$$

$$\overline{\lim_{t\to\infty}}\int_t^{t+T}\left\{\nu\|u_\varepsilon\|^2 + \frac{1}{\varepsilon}|\text{div } u_\varepsilon|^2\right\} ds \leqq \left(\frac{T}{\nu^2\lambda_1} + \frac{1}{\nu^3\lambda_1^2}\right)|f|^2. \tag{2.7}$$

Let ρ_0^2 denote the right-hand side of (2.5). It follows from (2.4) that if $|u_{\varepsilon 0}| \leqq \rho$, $\rho \geqq \rho_0$, then $|u_\varepsilon(t)| \leqq \rho$ for all $t > 0$, i.e., $S_\varepsilon(t)$ maps the ball of $\mathbb{L}^2(\Omega)$ centered at 0 of radius ρ into itself:

(2.8) For any $\rho \geqq \rho_0$ the ball of G centered at 0 of radius ρ is invariant for the semigroup $S_\varepsilon(t)$.

We note also that for $\rho > \rho_0$ any trajectory eventually enters into the ball of $\mathbb{L}^2(\Omega)$ centered at 0 of radius ρ. The time of entrance of $u_\varepsilon(t)$ into this ball is uniform for all the $u_{\varepsilon 0}$ such that

$$|u_{\varepsilon 0}| \leqq r_0\ (< \infty).$$

We recall that a set $\mathscr{C} \subset G$ is called an *absorbing set* for the semigroup $S_\varepsilon(t)$ if, for every bounded set $\mathscr{C}_0 \subset G$, there exists $t_0 = t_0(\mathscr{C}_0)$ such that

$$S_\varepsilon(t)\mathscr{C}_0 \subset \mathscr{C} \quad \forall t \geqq t_0(\mathscr{C}_0).$$

We then interpret the result above as follows:

(2.9) For any $\rho > \rho_0$, the ball of G centered at 0 of radius ρ is absorbing in G for the semigroup $S_\varepsilon(t)$.

The time of entrance of u_ε in this ball, depending on r_0, $\delta = \rho - \rho_0$ (and the data $\nu, \lambda_1, |f|$), $t = t_0(r_0, \delta)$ can be explicitly computed from (2.4) and we have

$$|u_\varepsilon(t)| \leqq \rho_0 + \delta \quad \text{for } t \geqq t_0(r_0, \delta),$$

$$\int_t^{t+T} \left(\|u_\varepsilon\|^2 + \frac{1}{\varepsilon}|\operatorname{div} u_\varepsilon|^2 \right) ds \leqq \frac{(\rho_0+\delta)^2}{\nu} + \frac{T}{\nu^2\lambda_1}|f|^2 \tag{2.10}$$

$$\forall T > 0, \quad \forall t \geqq t_0(r_0, \delta).$$

2.2. Uniform estimates in H^1. We now take the scalar product of (1.15) with $\mathscr{A}_\varepsilon u_\varepsilon$ in $\mathbb{L}^2(\Omega)$, the operator $\mathscr{A}_\varepsilon$ being defined in Lemma 1.1. We find

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left(\|u_\varepsilon\|^2 + \frac{1}{\varepsilon}|\operatorname{div} u_\varepsilon|^2 \right) + \nu|\mathscr{A}_\varepsilon u_\varepsilon|^2 \\
&\quad = (f - \tilde{\mathscr{B}}(u_\varepsilon), \mathscr{A}_\varepsilon u_\varepsilon) \\
&\quad \leqq (|f| + |\tilde{\mathscr{B}}(u_\varepsilon)|)|\mathscr{A}_\varepsilon u_\varepsilon| \\
&\quad \leqq \frac{\nu}{4}|\mathscr{A}_\varepsilon u_\varepsilon|^2 + \frac{1}{\nu}|f|^2 + |\tilde{\mathscr{B}}(u_\varepsilon)|\,|\mathscr{A}_\varepsilon u_\varepsilon|.
\end{aligned}$$

We know (cf. [6], [16]) that there exists a constant C_4 depending only on Ω such that

$$|\mathscr{B}(\varphi, \psi)| \leqq \begin{cases} C_4|\varphi|^{1/2}|\Delta\varphi|^{1/2}\|\psi\|, \\ C_4|\varphi|^{1/2}\|\varphi\|^{1/2}\|\psi\|^{1/2}|\Delta\psi|^{1/2} \end{cases} \quad \forall \varphi, \psi \in \mathbb{H}^2(\Omega). \tag{2.11}$$

It can be proved exactly in the same manner that

$$|\tilde{\mathscr{B}}(\varphi, \psi)| \leqq \begin{cases} C_5\{|\varphi|^{1/2}|\Delta\varphi|^{1/2}\|\psi\| + \|\varphi\|\,|\psi|^{1/2}|\Delta\psi|^{1/2}\}, \\ C_5\{|\varphi|^{1/2}\|\varphi\|^{1/2}\|\psi\|^{1/2}|\Delta\psi|^{1/2} + \|\varphi\|\,|\psi|^{1/2}|\Delta\psi|^{1/2}\} \end{cases} \quad \forall \varphi, \psi \in \mathbb{H}^2(\Omega). \tag{2.12}$$

Thus

$$|\tilde{\mathscr{B}}(u_\varepsilon)| = |\tilde{\mathscr{B}}(u_\varepsilon, u_\varepsilon)| \leqq 2C_5|u_\varepsilon|^{1/2}\|u_\varepsilon\|\,|\Delta u_\varepsilon|^{1/2}.$$

We then apply (1.23) (cf. Lemma 1.1 and Remark 1.2 for the space periodic case) and we bound this expression by

$$2C_5C_3^{1/2}|u_\varepsilon|^{1/2}\|u_\varepsilon\|\,|\mathscr{A}_\varepsilon u_\varepsilon|^{1/2},$$

and

$$\begin{aligned}
|(\tilde{\mathscr{B}}(u_\varepsilon), \mathscr{A}_\varepsilon u_\varepsilon)| &\leqq 2C_5C_3^{1/2}|u_\varepsilon|^{1/2}\|u_\varepsilon\|\,|\mathscr{A}_\varepsilon u_\varepsilon|^{3/2} \\
&\leqq \text{(with Young Inequality)} \\
&\leqq \frac{\nu}{4}|\mathscr{A}_\varepsilon u_\varepsilon|^2 + \frac{C_1'}{\nu^3}|u_\varepsilon|^2\|u_\varepsilon\|^4,
\end{aligned}$$

where C_1' as well as the C_j, C_j', $\cdots$, is a constant depending only on Ω.[4]

Combining these estimates we arrive at

$$\frac{d}{dt}\left(\|u_\varepsilon\|^2+\frac{1}{\varepsilon}|\operatorname{div} u_\varepsilon|^2\right)+\nu|\mathscr{A}_\varepsilon u_\varepsilon|^2\leqq\frac{2}{\nu}|f|^2+\frac{2C_1'}{\nu^3}|u_\varepsilon|\,\|u_\varepsilon\|^4. \tag{2.13}$$

First we deduce from (2.13) that

$$\begin{aligned} &y'\leqq h+gy,\\ &y=\|u_\varepsilon\|^2+\frac{1}{\varepsilon}|\operatorname{div} u_\varepsilon|^2,\quad h=\frac{2}{\nu}|f|^2,\\ &g=\frac{2C_1'}{\nu^3}|u_\varepsilon|\,\|u_\varepsilon\|^2. \end{aligned} \tag{2.14}$$

We apply the Uniform Gronwall Lemma [8] that we recall below for the convenience of the reader and we obtain that $y(t)$ is bounded uniformly with respect to t (and ε) on $[t_0+T,+\infty)$. More precisely

$$\|u_\varepsilon(t)\|^2+\frac{1}{\varepsilon}|\operatorname{div} u_\varepsilon(t)|^2\leqq\rho_1\quad\forall t\geqq t_0(r_0,\delta)+T\quad\forall T>0 \tag{2.15}$$

with (cf. (2.9) and Lemma 2.1 below)

$$\begin{aligned} &\rho_1=\left(\frac{\alpha_3}{T}+\alpha_2\right)\exp(\alpha_1),\\ &\alpha_1=\frac{2C_1'}{\nu^3}(\rho_0+\delta)\alpha_3,\quad\alpha_2=\frac{2}{\nu}|f|^2,\quad\alpha_3=\frac{(\rho_0+\delta)}{\nu}+\frac{T}{\nu^2\lambda_1}|f|^2. \end{aligned} \tag{2.16}$$

Then we obtain from (2.13)

$$\int_t^{t+T}|\mathscr{A}_\varepsilon u_\varepsilon|^2\,ds\leqq\frac{\rho_1(1+\alpha_1)}{\nu}+\frac{T\alpha_2}{\nu}\quad\forall T>0,\quad\forall t\geqq t_0(r_0,\delta), \tag{2.17}$$

$$\int_t^{t+T}|\Delta u_\varepsilon|^2\,ds\leqq\frac{C_3^2}{\nu}\{\rho_1(1+\alpha_1)+T\alpha_2\}\quad\forall T>0,\quad\forall t\geqq t_0(r_0,\delta). \tag{2.18}$$

We infer from (2.15), as in (2.8) that

(2.19) The ball of W centered at 0 of radius ρ_1 (given by (2.16)) is absorbing in W for the semigroup $S_\varepsilon(t)$.

Remark 2.1. The bounds similar to (2.15), for $0\leqq t\leqq t_0+T$, and to (2.17) for $0\leqq t\leqq t_0$ are easily derived from (2.14) using the classical Gronwall Lemma. Like in (2.15) and (2.17) the bounds are independent on ε.

Remark 2.2. We observe that in (2.8) and (2.19) the absorbing sets are *independent* of ε.

We conclude with the Uniform Gronwall Lemma.

LEMMA 2.1. *Let g, h, y be three positive locally integrable functions on $]t_0,\infty[$; assume that y is absolutely continuous and*

$$\frac{dy}{dt}\leqq gy+h\quad\textit{for } t\geqq t_0, \tag{2.20}$$

$$\int_t^{t+T}g(s)\,ds\leqq\alpha_1,\quad\int_t^{t+T}h(s)\,ds\leqq\alpha_2,\quad\int_t^{t+T}y(s)\,ds\leqq\alpha_3\quad\forall t\geqq t_0, \tag{2.21}$$

[4] More precisely, these constants depend on the shape of Ω but not on its size, i.e., they are the same for instance for Ω and $\lambda\Omega$ for all $\lambda>0$, or $\Omega+a$ for all $a\in\mathbb{R}^2$.

where T, $\alpha_1, \alpha_2, \alpha_3$ are positive constants. Then

$$y(t+T) \leqq \left(\frac{\alpha_3}{T} + \alpha_2\right) \exp(\alpha_1) \quad \forall t \geqq t_0. \tag{2.22}$$

Proof. Assume that $t_0 \leqq t \leqq s < t+T$. We deduce from (2.20) that

$$\frac{d}{ds}\left(y(s) \exp\left(-\int_t^s g(\tau)\, d\tau\right)\right) \leqq h(s) \exp\left(-\int_t^s g(\tau)\, d\tau\right) \leqq h(s).$$

Then by integration between s and $t+T$

$$y(t+T) \leqq y(s) \exp\left(\int_s^{t+T} g(\tau)\, d\tau\right) + \left(\int_s^{t+T} h(\tau)\, d\tau\right) \exp\left(\int_s^{t+T} g(\tau)\, d\tau\right).$$

Integration of this last inequality with respect to s between t and $t+T$ gives precisely (2.22). □

2.3. Existence of the universal attractors. We denote by $\mathscr{C}_0$ the $\mathbb{L}^2$-absorbing set introduced in (2.9), i.e., the ball of G centered at 0 of radius $\rho > \rho_0$, and we consider its ω-limit set for the semigroup S_ε:

$$X_\varepsilon = \bigcap_{s \geqq 0} \overline{\bigcup_{t \geqq s} S_\varepsilon(t)\mathscr{C}_0}, \tag{2.23}$$

where the closures are taken in $\mathbb{L}^2(\Omega)$. It is easy to see that $\varphi \in X_\varepsilon$ if and only if there exists a sequence $t_m \to \infty$ and a sequence φ_m of elements of $\mathscr{C}_0$ such that

$$S_\varepsilon(t_m)\varphi_m \to \varphi \quad \text{in } \mathbb{L}^2(\Omega) \quad \text{as } m \to \infty. \tag{2.24}$$

Exactly as in the case of the Navier-Stokes equations [6] or by application of general results [19], we can prove the following.

THEOREM 2.1. *The set X_ε is included and bounded in $\mathbb{H}^2(\Omega)$, compact in $\mathbb{L}^2(\Omega)$ and connected.*

This set is a functional invariant set for the semigroup S_ε, i.e.,

$$S_\varepsilon(t)X_\varepsilon = X_\varepsilon \quad \forall t > 0 \ (\textit{and thus } \forall t \in \mathbb{R}). \tag{2.25}$$

This set is an attractor in G and W. Its basin of attraction is the whole space G (resp. W). It is the largest bounded attractor and the largest bounded invariant set for S_ε.

This attractor X_ε is called the universal attractor for the semigroup S_ε. Our aim will be, in § 3, to study its convergence to the universal attractor for the Navier-Stokes equations.

2.4. Remark on another perturbed equation. As indicated in the Introduction, the nonlinear term $\frac{1}{2}(\operatorname{div} u_\varepsilon)u_\varepsilon$ was introduced in order to have a well-set Cauchy problem for arbitrary large data f and u_0 and arbitrary ε. Now since we are interested here in small values of ε, it is interesting to observe that we can also obtain global existence for the equation

$$\frac{\partial v_\varepsilon}{\partial t} - \nu\Delta v_\varepsilon + (v_\varepsilon \cdot \nabla)v_\varepsilon - \frac{\nu}{\varepsilon}\nabla(\operatorname{div} v_\varepsilon) = f$$

with the same initial and boundary conditions as those on u_ε, namely (0.4) or (0.5) and (0.7), provided ε is small enough. More precisely we claim that if $v_{\varepsilon_0} \in G$ and

$$|v_{\varepsilon_0}| \leqq R$$

(where R is an arbitrary real number), then for

$$0<\varepsilon\leqq \operatorname{Min}\left(\frac{1}{2R^2},\frac{\nu^2}{4|f|^2}\right)$$

the solution v_ε (with $v_\varepsilon(0)=v_{\varepsilon_0}$) exists and is unique in the spaces $L^2(0,T;W)\cap \mathscr{C}([0,T];G)$ for all $T>0$. Moreover (compare to (2.4))

$$|v_\varepsilon(t)|^2\leqq |v_{\varepsilon_0}|^2\, e^{-\nu\lambda_1 t/2}+\frac{2|f|^2}{\nu^2\lambda_1^2}(1-e^{-\nu\lambda_1 t/2}).$$

This inequality is proved as (2.4), but since $b(v_\varepsilon,v_\varepsilon,v_\varepsilon)=-\frac{1}{2}\int_\Omega(\operatorname{div} v_\varepsilon)|v_\varepsilon|^2\,dx$ we have, instead of (2.3),

$$\frac{1}{2}\frac{d}{dt}|v_\varepsilon|^2+\nu\|v_\varepsilon\|^2+\frac{2\nu}{\varepsilon}|\operatorname{div} v_\varepsilon|^2\leqq\frac{|f|^2}{\nu\lambda_1}+\frac{1}{2}\int_\Omega(\operatorname{div} v_\varepsilon)|v_\varepsilon|^2\,dx.$$

The supplementary term in the right-hand side is bounded by

$$\left|\frac{1}{2}\int_\Omega(\operatorname{div} v_\varepsilon)|v_\varepsilon|^2\,dx\right|\leqq\frac{\nu}{2\varepsilon}|\operatorname{div} v_\varepsilon|^2+\frac{\varepsilon}{8\nu}\int_\Omega|v_\varepsilon|^4\,dx.$$

Then using the interpolation inequality recalled hereafter in (3.48), we see that

$$\frac{d}{dt}|v_\varepsilon|^2+\nu\left(1-\frac{\varepsilon}{2}|v_\varepsilon|^2\right)\|v_\varepsilon\|^2\leqq\frac{|f|^2}{\nu\lambda_1}.$$

By the choice of ε, $1-(\varepsilon/2)|v_\varepsilon(t)|^2>\frac{1}{2}$ at time $t=0$, and by Gronwall's Lemma technics it follows that this holds true for all $t>0$. This guarantees the existence of $v_\varepsilon(t)$ for all time and provides a uniform estimate on $|v_\varepsilon(t)|$ for $t\geqq 0$ and ε as above.

If we choose

$$R\geqq\frac{\sqrt{2}|f|}{\nu\lambda_1}$$

then for every ε, $0<\varepsilon\leqq 1/2R^2$, the ball of G centered at 0 of radius R is invariant for the semigroup $\Sigma_\varepsilon(t)$ ($\Sigma_\varepsilon(t)v_{\varepsilon_0}\equiv v_\varepsilon(t)$). Moreover any ball of G centered at 0 of radius R_0, $R_0>|f|/\nu\lambda_1$, is an absorbing set for the semigroup $\Sigma_\varepsilon(t)$ in the ball of G centered at 0 and of radius R.

The uniform estimates in H^1 for u_ε (i.e., (2.15) and also (2.17), (2.18)) can also be derived on the solutions v_ε using similar methods. It follows then by the same procedure that Theorem 2.1 can be extended to this case: the semigroup Σ_ε possesses a universal attractor in the ball of G centered at 0 of radius R (where R is a fixed number greater than $\sqrt{2}|f|/\nu$ and ε satisfies $0<\varepsilon\leqq 1/2R^2$).

3. Convergence to 0 of the penalty parameter. We recall a result on the convergence of attractors (§ 3.1). Then we recall some known results on the convergence of u_ε to u as $\varepsilon\to 0$ and give some complements (§ 3.2). Finally in § 3.3 we apply the results of §§ 3.1 and 3.2 and establish the convergence of the universal attractor of the penalized Navier-Stokes equations to that of the exact N.S.E.

3.1. A result on convergence of attractors. We give a result on convergence of attractors; for another form of this result see Foias and Temam [7]. Although the notations are the same as in the rest of the article, this § 3.1 is self-contained and independent.

We are given a Hilbert space G and a Hilbert subspace $H \subset G$ and we denote by P the projector in G onto H. We consider in H a semigroup of operators $\{S(t)\}_{t\geq 0}$ which possesses an attractor $\mathscr{A}$ attracting the whole space H.

We are also given a family of semigroups $\{S_\varepsilon(t)\}_{t\geq 0}$ which depends on a parameter ε, $0<\varepsilon\leq\varepsilon_0$. For each ε it is assumed that S_ε possesses an attractor $\mathscr{A}_\varepsilon$ which attracts the whole space G. Furthermore the following assumptions are made:

(3.1) There exists a fixed open bounded set of G, $\mathscr{U}$, which contains $\mathscr{A}$ and $\mathscr{A}_\varepsilon$ $\forall\varepsilon$, and is invariant by $S_\varepsilon(t)$ and $S(t)P$, ($S_\varepsilon(t)\mathscr{U}\subset\mathscr{U}$, $S(t)P\mathscr{U}\subset\mathscr{U}$, $\forall\varepsilon$, $\forall t>0$).

(3.2) For every bounded set $\mathscr{C}\subset G$ and every $t\in]0,\infty[$

$$\mathrm{Sup}_{u_1\in\mathscr{C}}\,|(I-P)S_\varepsilon(t)u_1|\to 0 \text{ as } \varepsilon\to 0.$$

(3.3) For every compact $I\subset]0,\infty[$

$$\mathrm{Sup}_{t\in I}\,|S_\varepsilon(t)v_1-S(t)Pv_1|\leq a(|v_1-Pv_1|)+b(\varepsilon)$$

uniformly with respect to v_1 in a compact of G, where $a(\alpha)$, $b(\alpha)$ are continuous increasing functions which tend to 0 as $\alpha\to 0$.[5]

PROPOSITION 3.1. *Under the above assumptions $\mathscr{A}_\varepsilon$ converges to $\mathscr{A}$ as $\varepsilon\to 0$ in the following sense*:

(3.4) *For every open neighborhood $\mathscr{V}$ of $\mathscr{A}$ there exists ε_1 depending on $\mathscr{V}$ and for $\varepsilon\leq\varepsilon_1$, $\mathscr{A}_\varepsilon\subset\mathscr{V}$.*

In particular $(I-P)\,\mathscr{A}_\varepsilon\to 0$ (in the above sense) as $\varepsilon\to 0$.

Proof. Let $\mathscr{V}_\alpha(\mathscr{A})$ denote the α-neighborhood of $\mathscr{A}$ in H, i.e., the union of open balls of radius α of H centered on $\mathscr{A}$. It is sufficient to show that for every α, there exists $\varepsilon_1=\varepsilon_1(\alpha)$ and $t_1=t_1(\alpha)$ such that, for $0<\varepsilon\leq\varepsilon_1(\alpha)$ and $t\geq t_1(\alpha)$,

$$S_\varepsilon(t)\mathscr{U}\subset\mathscr{V}_\alpha(\mathscr{A}). \tag{3.5}$$

Indeed the ω-limit set of $\mathscr{U}$ for S_ε, $\omega_\varepsilon(\mathscr{U})$ is equal to $\mathscr{A}_\varepsilon$[6] and (3.5) implies that

$$\omega_\varepsilon(\mathscr{U})=\mathscr{A}_\varepsilon\subset\mathscr{V}_\alpha(\mathscr{A}).$$

We now prove (3.5); there exists $r_0>0$ such that $(\mathscr{U}\supset)P\mathscr{U}\supset\mathscr{V}_{r_0}(\mathscr{A})$. Assuming that $\alpha>r_0$, and since $\mathscr{A}$ attracts $P\mathscr{U}$, we can find $t_1=t_1(\alpha)$ such that for $t\geq t_1/2$

$$S(t)P\mathscr{U}\subset\mathscr{V}_{\alpha/2}(\mathscr{A}). \tag{3.6}$$

We use (3.2) with $\mathscr{C}=\mathscr{U}$ and $t=t_1/2$ and we find that

$$\mathrm{Sup}_{u_0\in\mathscr{U}}\left|(I-P)S_\varepsilon\left(\frac{t_1}{2}\right)u_0\right|=\delta(\varepsilon)\to 0 \quad \text{as } \varepsilon\to 0. \tag{3.7}$$

We then use (3.3) with $I=[\frac{1}{2}t_1,\frac{3}{2}t_1]$, and $v_1=S_\varepsilon(\frac{1}{2}t_1)u_1$, $u_1\in\mathscr{U}$ and we obtain (note that $S(t_1/2)\mathscr{U}$ is a compact of G)

$$\mathrm{Sup}_{(1/2)t_1\leq t\leq(3/2)t_1}\,|S_\varepsilon(t)v_1-S(t)Pv_1|\leq a(|v_1-Pv_1|)+b(\varepsilon)\leq a(\delta(\varepsilon))+b(\varepsilon). \tag{3.8}$$

We can find $\varepsilon_1=\varepsilon_1(\alpha)$ such that $a(\delta(\varepsilon))+b(\varepsilon)\leq\alpha/2$ for every $\varepsilon\leq\varepsilon_1(\alpha)$. Then (3.6), (3.7) and (3.8) (with $v_1=S_\varepsilon(\frac{1}{2}t_1)u_1$, $u_1\in\mathscr{U}$) show that

$$S_\varepsilon(t)u_1\in\mathscr{V}_\alpha(\mathscr{A}) \quad \forall u_1\in\mathscr{U} \quad \forall t\in[t_1,2t_1] \quad \forall\varepsilon\leq\varepsilon_1(\alpha).$$

We have thus proved (3.5) for $t\in[t_1,2t_1]$.

---

[5] If (3.3) is satisfied with functions a, b which are not increasing, we can replace these functions by $\mathrm{Sup}_{0\leq r\leq\alpha}\,a(r)$, $\mathrm{Sup}_{0\leq r\leq\alpha}\,b(r)$ and (3.3) will be satisfied.

[6] $\mathscr{A}_\varepsilon=\omega_\varepsilon(\mathscr{A}_\varepsilon)\subset\omega_\varepsilon(\mathscr{U})$ since $\mathscr{A}_\varepsilon\subset\mathscr{U}$ and $\omega_\varepsilon(\mathscr{U})\subset\mathscr{A}_\varepsilon$ since $\mathscr{A}_\varepsilon$ attracts $\mathscr{U}$.

In order to prove (3.5) for $t > 2t_1$ we write such a t in the form $t = nt_1 + \tau$ for an appropriate integer n and $\tau \in [t_1, 2t_1]$. Then if $u_1 \in \mathcal{U}$,

$$S_\varepsilon(t)u_1 = S_\varepsilon(\tau)S_\varepsilon(nt_1)u_1. \tag{3.9}$$

Due to (3.1), $S_\varepsilon(nt_1)u_1 \in \mathcal{U}$ and (3.5) (which is valid for $\tau \in [t_1, 2t_1]$) implies that $S_\varepsilon(t)u_1 \in \mathcal{V}_\alpha(\mathcal{A})$, if $\varepsilon \leqq \varepsilon_1(\alpha)$.

The proposition is proved. □

This result will be applied to the Navier-Stokes equations, but before that we study the convergence of u_ε to u as $\varepsilon \to 0$.

3.2. Convergence of the solutions of the penalized problems. It was proved in [13], [14] that, if $u_\varepsilon(0) = u(0) \in H$ then u_ε converges to u_ε as $\varepsilon \to 0$, in the following sense:

$$u_\varepsilon \to u \quad \text{in } L^2(0, T; \mathbb{H}^1(\Omega)) \quad \text{and} \quad \mathscr{C}([0, T]; \mathbb{L}^2(\Omega)) \qquad \forall T < \infty. \tag{3.10}$$

Here we will improve this result and, in view of (3.2), (3.3), consider also the case where $u_\varepsilon(0) \notin H$.

We investigate the case where $u_\varepsilon(0)$—hereafter denoted u_1—is fixed and belongs to G, not necessarily to H. We use the a priori estimates of § 2 which are of course valid. We need only these estimates on a finite interval $[0, T]$. We infer from (2.4), (2.6) that

$$u_\varepsilon \text{ is bounded in } L^\infty(0, T; G) \cap L^2(0, T; W) \text{ independently of } \varepsilon, \tag{3.11}$$

$$\frac{1}{\sqrt{\varepsilon}} \operatorname{div} u_\varepsilon \text{ is bounded in } L^2(0, T; L^2(\Omega)) \text{ independently of } \varepsilon. \tag{3.12}$$

Then we use (2.13), (2.14) but in a slightly different manner than in § 2. We infer from (2.14) that

$$(ty)' \leqq th + (1 + tg)y = th + y + tgy. \tag{3.13}$$

Then by the classical Gronwall Lemma,

$$\begin{aligned} ty(t) &\leqq \left(\int_0^t sh(s) + y(s))\, ds \right) \exp\left(\int_0^t g(s)\, ds \right) \\ &\leqq \left(T \int_0^T h(s)\, ds + \int_0^T y(s)\, ds \right) \exp\left(\int_0^T g(s)\, ds \right). \end{aligned}$$

Due to (3.11), (3.12) and the expressions of y, g, h in (2.14), we conclude that

$$\|u_\varepsilon(t)\|^2 + \frac{1}{\varepsilon} |\operatorname{div} u_\varepsilon(t)|^2 \leqq \frac{K_1}{t} \quad \forall t \in [0, T] \tag{3.14}$$

where K_1 depends on the data (ν, f, u_1) but not on ε and t. Thus using again (2.13) in conjunction with (1.23) we obtain

$$\int_0^T t |\mathcal{A}_\varepsilon u_\varepsilon(t)|^2\, dt \leqq K_2, \tag{3.15}$$

$$\int_0^T t |\Delta u_\varepsilon(t)|^2\, dt \leqq K_2' \tag{3.16}$$

where K_2, K_2' depend on the data but not on ε.

Then we consider (1.15), which is valid in G for almost every t in $[0, T]$ and we project it on H. We find

$$\frac{d}{dt} P u_\varepsilon - \nu P \Delta u_\varepsilon + P\tilde{\mathcal{B}}(u_\varepsilon) = Pf. \tag{3.17}$$

The term involving grad div u_ε has been annihilated by P. With (2.15), (3.14), (3.16) we obtain

$$\frac{d}{dt}(tPu_\varepsilon) \text{ is bounded in } L^2(0, T; H) \text{ independently of } \varepsilon. \tag{3.18}$$

Because of (3.14), (3.16) and since P is continuous in $\mathbb{H}^m(\Omega)$ for all $m \in \mathbb{N}$ (cf. [15, Chap. I, Remark 1.6]), we have

$$tPu_\varepsilon \text{ is bounded in } L^2(0, T; \mathbb{H}^2(\Omega)) \cap L^\infty(0, T; \mathbb{H}^1(\Omega)) \text{ independently of } \varepsilon. \tag{3.19}$$

It follows from a compactness theorem in [15, Chap. III, § 2] that tPu_ε is relatively compact in $L^2(0, T; \mathbb{H}^1(\Omega))$; hence there exists a subsequence, still denoted ε, which tends to 0 such that

$$u_\varepsilon \to u \quad \text{in } L^2(0, T; \mathbb{H}^1(\Omega)) \quad \text{weakly}, \tag{3.20}$$

$$tPu \to tPu \quad \text{in } L^2(0, T; \mathbb{H}^1(\Omega)) \quad \text{strongly}, \tag{3.21}$$

$$\text{div } u_\varepsilon \to 0 \quad \text{in } L^2(0, T; \mathbb{L}^2(\Omega)) \quad \text{strongly}. \tag{3.22}$$

The term $(I - P)u_\varepsilon$ is the gradient of a function q_ε such that (cf. [15, Chap. I, Remark 1.6])

$$\begin{aligned} &\Delta q_\varepsilon = \text{div } u_\varepsilon \quad \text{in } \Omega, \\ &\frac{\partial q_\varepsilon}{\partial \nu} = 0 \quad \text{on } \Gamma \quad \text{in the case (0.4)}, \\ &q_\varepsilon \text{ is periodic in the case (0.5)}. \end{aligned} \tag{3.23}$$

Hence

$$|(I - P)u_\varepsilon|_{\mathbb{H}^1(\Omega)} \leqq |\text{grad } q_\varepsilon|_{\mathbb{H}^1(\Omega)} \leqq C_7 |\text{div } u_\varepsilon| \tag{3.24}$$

and due to (3.12), (3.21)

$$(I - P)u_\varepsilon \to 0 \quad \text{in } L^2(0, T; \mathbb{H}^1(\Omega)), \tag{3.25}$$

$$tu_\varepsilon \to tu \quad \text{in } L^2(0, T; \mathbb{H}^1(\Omega)) \quad \text{strongly}. \tag{3.26}$$

It is clear that u belongs to $L^2(0, T, V) \cap L^\infty(0, T; H)$ ($Pu = u$). The passage to the limit in (3.17) can be made by standard methods (cf. [15]) and we find at the limit that u satisfies (1.13). We can also deduce from (3.17) and (3.11) that $(d/dt)Pu_\varepsilon$ is bounded in $L^2(0, T; V')$ independently of ε. This implies that $u' \in L^2(0, T; V')$, $u \in \mathscr{C}([0, T]; H)$ and that

$$Pu_\varepsilon(t) \to u(t) \quad \text{as } \varepsilon \to 0 \text{ weakly in } V' \quad \forall t \in [0, T]. \tag{3.27}$$

In particular $u(0) = Pu_1$. We conclude that u is the unique solution of (1.13) satisfying $u(0) = Pu_1$ and that all the convergences above are valid for the whole sequence $\varepsilon \to 0$ and not only for a subsequence. We have thus proved the following proposition.

PROPOSITION 3.2. *When $\varepsilon \to 0$, the solution u_ε of* (1.15), (1.16) *such that $u_\varepsilon(0) = u_1 \in G$, converges to the solution u of* (1.13), (1.14) *such that $u(0) = u_0 = Pu_1$. The convergences hold in particular in the sense* (3.20), (3.25), (3.26).

If $u_1 \in G \setminus H$, $u_1 \neq u_0$, $u_\varepsilon(0)$ does not converge to $u(0)$ and a boundary layer appears near $t = 0$. However away from $t = 0$, on a compact subset of $]0, T]$, one can improve the convergence of u_ε to u; this will be discussed hereafter.

Before that we observe that (1.13) possesses a regularizing property, and although $u_0 \in H$, we have $u(t) \in V$ for all $t > 0$ (since $f \in H$). A classical energy inequality for (1.13) reads (cf. [16])

$$\frac{d}{dt}\|u\|^2 + \nu|Au|^2 \leqq \frac{2}{\nu}|f|^2 + \frac{c_1'}{\nu^3}|u|^2\|u\|^4 \tag{3.28}$$

where c_1' is an appropriate constant and after multiplication by t we obtain

$$\frac{d}{dt}(t\|u\|^2) + \nu t|Au|^2 \leqq \|u\|^2 + \frac{2t}{\nu}|f|^2 + \frac{c_1' t}{\nu^3}|u|^2\|u\|^4. \tag{3.29}$$

Thanks to the Gronwall inequality and the fact already known that $u \in L^2(0, T; V) \cap L^\infty(0, T; H)$ we infer from (3.29) that

$$\sqrt{t}u \in L^\infty(0, T; V) \cap L^2(0, T; D(A)). \tag{3.30}$$

After multiplication of (1.13) by t we see that

$$\frac{d}{dt}(tu) + \nu tAu + tB(u) = u + tPf \tag{3.31}$$

and we conclude from (3.30) and (2.11) that

$$\frac{d}{dt}(tu) \in L^2(0, T; H) \subset L^2(0, T; \mathbb{L}^2(\Omega)). \tag{3.32}$$

Considering then the Navier-Stokes equation itself, i.e., (0.1), we see after multiplication by t and utilization of (2.11), (3.30), (3.32) that

$$\text{grad}\,(tp) \in L^2(0, T; \mathbb{L}^2(\Omega)), \tag{3.33}$$

which implies

$$tp \in L^2(0, T; H^1(\Omega)). \tag{3.34}$$

We now improve the convergence of u_ε to u. We subtract (0.1) from (0.3) and set $v_\varepsilon = u_\varepsilon - u$, $q_\varepsilon = -\nu/\varepsilon \text{ div } u_\varepsilon - p$. We obtain

$$\frac{\partial v_\varepsilon}{\partial t} - \nu\Delta v_\varepsilon + \tilde{\mathcal{B}}(u_\varepsilon, v_\varepsilon) + \tilde{\mathcal{B}}(v_\varepsilon, u) + \text{grad}\, q_\varepsilon = 0, \tag{3.35}$$

and we then take the scalar product of this equation with v_ε in $\mathbb{L}^2(\Omega)$:

$$\frac{1}{2}\frac{d}{dt}|v_\varepsilon|^2 + \nu\|v_\varepsilon\|^2 + (\tilde{\mathcal{B}}(v_\varepsilon, u), v_\varepsilon) + (\text{grad}\, q_\varepsilon, v) = 0. \tag{3.36}$$

The first term involving $\tilde{\mathcal{B}}$ has disappeared due to (2.1) while the remaining term is majorized using the following inequality:

$$|(\tilde{\mathcal{B}}(\varphi, \psi), \theta| \leqq c_6\{|\psi|^{1/2}\|\varphi\|^{1/2}\|\psi\| + \|\varphi\||\psi|^{1/2}\|\psi\|^{1/2}\}|\theta|^{1/2}\|\theta\|^{1/2}. \tag{3.37}$$

This inequality is proved exactly as the similar inequality for $\tilde{\mathcal{B}}$ (cf. Remark 2.2 in [16]). It implies

$$\begin{aligned}|(\tilde{\mathcal{B}}(v_\varepsilon, u), v_\varepsilon)| &\leqq c_6\{|v_\varepsilon|\,\|v_\varepsilon\|\,\|u\| + |v_\varepsilon|^{1/2}\|v_\varepsilon\|^{3/2}|u|^{1/2}\|u\|^{1/2}\} \\ &\leqq \text{(with Schwarz and Young inequalities)} \\ &\leqq \frac{\nu}{2}\|v_\varepsilon\|^2 + \frac{c_2'}{\nu}\left(1 + \frac{1}{\nu^2}|u|^2\right)\|u\|^2|v_\varepsilon|^2.\end{aligned}$$

We have

$$(\text{grad } q_\varepsilon, v_\varepsilon) = -(q_\varepsilon, \text{div } v_\varepsilon)$$

$$= \frac{\nu}{\varepsilon}|\text{div } u_\varepsilon|^2 + (p, \text{div } u_\varepsilon)$$

$$\geqq \frac{\nu}{2\varepsilon}|\text{div } u_\varepsilon|^2 - \frac{\varepsilon}{2\nu}|p|^2.$$

Hence

$$\frac{d}{dt}|v_\varepsilon|^2 + \nu\|v_\varepsilon\|^2 + \frac{\nu}{\varepsilon}|\text{div } u_\varepsilon|^2 \leqq \frac{\varepsilon}{\nu}|p|^2 + g|v_\varepsilon|^2, \tag{3.38}$$

$$g = \frac{2c_2'}{\nu}\left(1 + \frac{1}{\nu^2}|u|^2\right)\|u\|^2. \tag{3.39}$$

We are going to show that (3.38) implies the following property which will allow us to prove (3.3):

(3.40) On every compact interval $I = [t_0, T]$, $0 < t_0 < T < \infty$, and for every $R > 0$, $\lim_{\varepsilon \to 0} \text{Sup}_{|u_1| \leqq R, t \in I} |S_\varepsilon(t)u_1 - S(t)Pu_1| = 0$.

The proof of this property relies on (3.38) and the fact that for every $t_0 > 0$, we have

$$\lim_{\varepsilon \to 0} \sup_{|u_1| \leqq R} |S_\varepsilon(t_0)u_1 - S(t_0)Pu_1| = 0. \tag{3.41}$$

Indeed, let us assume for the moment that (3.41) holds true. We notice that when u_1 belongs to G and $|u_1| \leqq R$, we have $|Pu_1| \leqq R$ and then since $u(t) = S(t)Pu_1$ satisfies the relation similar to (2.2)

$$\frac{d}{dt}|u|^2 + \nu\|u\|^2 \leqq \frac{|f|^2}{\nu\lambda_1},$$

it follows that there exists C_3 depending only on R, T, f, ν and Ω such that the norm of g (see (3.39)) in $L^1(0, T)$ is bounded by C_3. The same holds for the norm of tp in $L^2(0, T; H^1(\Omega))$ (see (3.34)) and we have ($|\varphi| \leqq \|\varphi\|_{H^1}$)

$$\text{Sup}_{|u_1| \leqq R} \int_0^T (t^2|p(t)|^2 + g(t))\, dt \leqq C_3 < \infty. \tag{3.42}$$

We return to (3.38) which implies

$$\frac{d}{dt}\left(|v_\varepsilon(t)|^2 \exp\left(-\int_0^t g(s)\, ds\right)\right) \leqq \frac{\varepsilon}{\nu}|p(t)|^2 \exp\left(-\int_0^t g(s)\, ds\right).$$

By integration between t_0 and t, we find

$$|v_\varepsilon(t)|^2 \leqq e^{C_3}\left\{|v_\varepsilon(t_0)|^2 + \frac{\varepsilon}{\nu}\int_{t_0}^t |p(s)|^2\, ds\right\} \quad \text{for } t_0 \leqq t \leqq T;$$

thus

$$|v_\varepsilon(t)|^2 \leqq e^{C_3}\left\{|v_\varepsilon(t_0)|^2 + \frac{\varepsilon C_3}{\nu t_0^2}\right\}, \qquad t_0 \leqq t \leqq T.$$

Keeping in mind that $v_\varepsilon(t) = S_\varepsilon(t)u_1 - S(t)Pu_1$, we see that (3.41) and this last estimate prove (3.40).

We proceed to the proof of (3.41). We argue by contradiction and assume that there exist $\delta > 0$ and a sequence $(\varepsilon_j, u_{1j}) \in \mathbb{R}_+^* \times G$ such that $|u_{1j}| \leqq R$, $\varepsilon_j \to 0$ as $j \to \infty$ and

$$|S_{\varepsilon j}(t_0)u_{1j} - S(t_0)Pu_{1j}| \geqq \delta > 0. \tag{3.43}$$

Since u_{1j} is bounded in G we can extract a weakly convergent subsequence, still denoted u_{1j}, such that $u_{1j} \to u_1$ in the weak topology of G. It can be shown with the methods used above that $S_{\varepsilon j}(\cdot)u_{1j}$ and $S(\cdot)Pu_{1j}$ converge to $S(\cdot)P_{u_1}$ for various topologies. In particular it follows from (2.15) that $S_{\varepsilon j}(t_0)u_{1j}$ is bounded in $\mathbb{H}^1(\Omega)$ and as in Proposition 3.2 it can be shown by compactness that $S_{\varepsilon j}(t_0)u_{1j}$ converges to $S(t_0)Pu_1$ in the norm of G, as $j \to \infty$. Similarly, $S(t_0)Pu_{1j}$ converges to $S(t_0)Pu_1$ in the norm of $G(H)$ as $j \to \infty$. These convergences contradict (3.43) and therefore (3.41) follows.

Remark 3.1. We have used (3.42), which is valid for the two boundary conditions (0.4) and (0.5). In fact, in the space periodic case, it can be proved that

$$\operatorname*{Sup}_{|v_1| \leqq R} \int_0^T (|p(t)|^2 + g(t))\, dt < \infty,$$

which is a stronger result and permits us to avoid the derivation of (3.34).

3.3. Convergence of the universal attractors. We now apply Proposition 3.1 to the convergence of the penalized attractors. The spaces G, H, the semigroups S, S_ε, will be the same as in the rest of the article. We must verify the assumptions (3.1) and (3.2). With (2.5), (2.8), (2.9) we see that any ball of G, centered at 0 of radius $\rho > \rho_0 = (1/\nu\lambda_1)|f|$ contains the attractor $\mathscr{A}_\varepsilon$ for all $\varepsilon > 0$, and is mapped into itself by $S_\varepsilon(t)$. It is proved (see for instance [18]) exactly as for (2.5) that

$$\overline{\lim_{t \to \infty}} |u(t)|^2 \leqq \frac{1}{\nu^2 \lambda_1^2} |f|^2,$$

when u is any solution of (1.13) (the exact Navier-Stokes equations) and thus $\mathscr{A}$ is contained in the ball of H centered at 0 of radius ρ. This shows that (3.1) is satisfied if we take $\mathscr{U}$ = the ball of G centered at 0 of radius $\rho > \rho_0$.

The proof of (3.2) follows from (3.14) and (3.23), (3.24):

$$|(I-P)S_\varepsilon(t)u_1| = |(I-P)u_\varepsilon(t)| \leqq C_7 |\operatorname{div} u_\varepsilon(t)| \leqq C_7 \sqrt{\frac{K_1 \varepsilon}{t}}.$$

It can be easily seen that K_1, as well as all the bounds in the estimates (3.11)-(3.16) depend only on r_1 when $u_1 \in G$ and $|u_1| \leqq r_1$, and not explicitly on u_1[7]; (3.2) is proved. Finally (3.3) follows readily from (3.40).

By application of Proposition 3.1 we obtain the following theorem.

THEOREM 3.1. *Let $\mathscr{A}_\varepsilon$ denote the universal attractor for the penalized Navier-Stokes equations and let $\mathscr{A}$ denote the universal attractor for the exact Navier-Stokes equations (in space dimension 2).*

Then, as $\varepsilon \to 0$, $\mathscr{A}_\varepsilon$ converges to $\mathscr{A}$ in the sense of (3.4) *and in particular $(I-P)\mathscr{A}_\varepsilon$ converges to* 0.

3.4. Remark on the dimensions. It was proved (Foiaş and Temam [6]) that the universal attractor $\mathscr{A}$ describing the long-time behavior of the exact N.S.E. is finite

[7] Everything follows from (2.4), where we write $|u_\varepsilon(0)|^2 = |u_1|^2 \leqq r_1^2$.

dimensional. In the two-dimensional case ([17]) the fractal[8] (and Hausdorff) dimension of $\mathscr{A}$ is $\leqq c_0(1+G)$, where c_0 is a universal constant and G is the nondimensional (Grashof) number

$$G=\frac{|f|}{\nu^2\lambda_1}. \tag{3.44}$$

In this section we apply a general result of Constantin, Foias and Temam [4] to the penalized N.S.E. and estimate in a similar fashion the dimension of $\mathscr{A}_\varepsilon$. We have the following theorem.

THEOREM 3.2. *The fractal (and Hausdorff) dimension of the universal attractor $\mathscr{A}_\varepsilon$ is $\leqq c_0(1+G+c_1\varepsilon G^2)$, where c_0 and c_1 are universal constants.*

The proof of this result is parallel to the one performed on Navier-Stokes equations in [4], [17], [18]. We just mention here the step of the proof which provides the result on the dimensions. For simplicity we consider the boundary condition (0.4). The space periodic case is totally similar. For $m\in\mathbb{N}^*$ we introduce a family $\{v^i\}_{i=1}^m$ in $\mathbb{H}_0^1(\Omega)$ which is *orthonormal* (o.n.) *in* $\mathbb{L}^2(\Omega)$. Let u_{ε_0} be an arbitrary point in $\mathscr{A}_\varepsilon$ and $\{u_\varepsilon\}$: $u_\varepsilon(t)=S_\varepsilon(t)u_{\varepsilon_0}$, be the solution of (1.15), (1.16). We form the quantity[9]

$$\sigma_m(u_{\varepsilon_0},\{v^i\},t)=\sum_{i=1}^m\left\{\nu\|v^i\|^2+\frac{\nu}{\varepsilon}|\operatorname{div} v^i|^2+\tilde{b}(v^i,u_\varepsilon(t),v^i)\right\} \tag{3.45}$$

and we set

$$q_m=\limsup_{t\to+\infty}\ \sup_{u_{\varepsilon_0}\in\mathscr{A}_\varepsilon}\frac{1}{t}\int_0^t\ \sup_{\{v^i\}_{i=1}^m \text{ o.n. in } \mathbb{L}^2}(-\sigma(u_{\varepsilon_0},\{v^i\},s))\,ds. \tag{3.46}$$

According to [4], if there exists some $m_0\in\mathbb{N}^*$ such that

$$q_{m_0}>0 \tag{3.47}$$

then $\mathscr{A}_\varepsilon$ has a finite fractal (and Hausdorff) dimension and

$$d_F(\mathscr{A}_\varepsilon)\leqq m_0\max_{1\leqq l\leqq m_0}\left(1-\frac{|q_l|}{q_{m_0}}\right). \tag{3.48}$$

Our aim now is to estimate (3.45) in order to obtain (3.47) for some $m_0\in\mathbb{N}^*$. We notice that, pointwise,

$$\left|\frac{1}{2}(\operatorname{div} v^i)u\cdot v^i\right|\leqq\frac{\nu}{\varepsilon}(\operatorname{div} v^i)^2+\frac{\varepsilon}{16\nu}|u|^2|v^i|^2; \tag{3.49}$$

therefore if we set

$$\rho(x)=\sum_{i=1}^m|v^i(x)|^2 \tag{3.50}$$

then thanks to successive applications of the Schwarz inequality we find that

$$|\tilde{b}(v_i,u_\varepsilon,v_i)|\leqq\|u_\varepsilon\|\,|\rho|+\frac{\nu}{\varepsilon}|\operatorname{div} v^i|^2+\frac{\varepsilon}{16\nu}\left(\int_\Omega|u_\varepsilon|^4\,dx\right)^{1/2}|\rho|. \tag{3.51}$$

[8] If $\mathscr{E}$ is a metric space, its fractal dimension (or capacity) $d_F(\mathscr{E})$ is defined as the lim sup as $r\to0^+$ of the ratio $N_2(\mathscr{E})/\log(1/r)$, where $N_r(\mathscr{E})$ is the minimum number of balls of radius r necessary to cover $\mathscr{E}$. The fractal dimension is always greater or equal to the Hausdorff dimension, the converse being false in general.

[9] This quantity occurs naturally when one considers the linearized flow around the trajectory $\{u_\varepsilon(t)\}$ on the attractor.

Returning to (3.45) we find that

$$\sigma_m(t) \geqq \nu\left(\sum_{i=1}^{m} |v^i\|^2\right) - \left(\|u_\varepsilon\| + \frac{\varepsilon}{16\nu}\left(\int_\Omega |u_\varepsilon|^4\, dx\right)^{1/2} |\rho|\right). \tag{3.52}$$

We set

$$\beta_\varepsilon \equiv \limsup_{t\to+\infty} \operatorname*{Sup}_{u_{\varepsilon_0}\in\mathscr{A}_\varepsilon} \left\{\frac{1}{t}\int_0^t \left(\|u_\varepsilon(s)\| + \frac{\varepsilon}{16\nu}\left(\int_\Omega |u_\varepsilon|^4\, dx\right)^{1/2}\right) ds\right\} \tag{3.53}$$

and then according to (3.52), (3.46) yields

$$q_m \geqq \nu\left(\sum_{i=1}^{m} \|v^i\|^2\right) - \beta_\varepsilon|\rho|. \tag{3.54}$$

We are going to use now the Lieb-Thirring inequality [9] (see also [19]) which is an improvement of the classical Gagliardo-Nirenberg-Sobolev inequality (see for instance [15, p. 291]):

$$\int_\Omega |\phi|^4\, dx \leqq 2|\phi|^2 \|\phi\|^2 \quad \forall \phi \in \mathbb{H}_0^1(\Omega). \tag{3.55}$$

According to the Lieb-Thirring inequality there exists a universal constant K (which is in particular independent of m) such that[10]

$$\int_\Omega \left(\sum_{i=1}^{m} |v^i(x)|^2\right)^2 dx \leqq K \sum_{i=1}^{m} \int_\Omega |\nabla v^i(x)|^2\, dx. \tag{3.56}$$

From (3.54) and (3.56), it follows that

$$q_m \geqq \frac{\nu}{K}|\rho|^2 - \beta_\varepsilon|\rho| \geqq \frac{\nu}{2K}|\rho|^2 - \frac{K}{2\nu}\beta_\varepsilon^2,$$

and since the $\{v^i\}$ are orthonormal in $\mathbb{L}^2(\Omega)$,

$$m = \int_\Omega \rho(x)\, dx \leqq |\rho||\Omega|^{1/2}$$

where $|\Omega|$ denotes the area of Ω. Hence

$$q_m \geqq \frac{\nu}{2K}\frac{m^2}{|\Omega|} - \frac{K}{2\nu}\beta_\varepsilon^2, \qquad m \geqq 1. \tag{3.57}$$

If we define m_0 by

$$m_0 - 1 < \frac{2K}{\nu}|\Omega|^{1/2}\beta_\varepsilon \leqq m_0, \tag{3.58}$$

then

$$-\frac{q_l}{q_{m_0}} \leqq 1 \quad \text{for } l = 1, \cdots, m_0 - 1$$

and from (3.48), it follows that

$$d_F(\mathscr{A}_\varepsilon) \leqq 2\left(1 + \frac{2K\beta_\varepsilon|\Omega|^{1/2}}{\nu}\right). \tag{3.59}$$

[10] Applying (3.55) to each of the v^i provides (3.56) with only a constant $K = K(m)$ and $K(m) \to \infty$ when $m \to \infty$.

It remains to estimate β_ε. According to (2.5) we have, for every $u_{\varepsilon_0} \in \mathscr{A}_\varepsilon$,

$$\limsup_{t\to+\infty} |u_\varepsilon(t)|^2 \leqq \frac{|f|^2}{\nu^2\lambda_1^2}, \tag{3.60}$$

and returning to (2.2) it follows by integration that

$$\limsup_{t\to+\infty} \frac{1}{t}\int_0^t \|u_\varepsilon(t)\|^2\, ds \leqq \frac{|f|^2}{\nu^2\lambda_1}. \tag{3.61}$$

Thanks to (3.55), we have

$$\int_\Omega |u_\varepsilon(x,t)|^4\, dx \leqq 2|u_\varepsilon(t)|^2\|u_\varepsilon(t)\|^2$$

and by (3.60), (3.61),

$$\limsup_{t\to+\infty} \frac{1}{t}\int_0^t \left(\int_\Omega |u_\varepsilon(x,s)|^4\, dx\right) ds \leqq 2\frac{|f|^2}{\nu^4\lambda_1^3}. \tag{3.62}$$

Now from (3.53) and (3.61), (3.62) it follows that

$$\beta_\varepsilon \leqq \frac{|f|}{\nu\lambda_1^{1/2}} + \frac{\sqrt{2}\varepsilon}{16}\frac{|f|^2}{\nu^3\lambda_1^{3/2}}. \tag{3.63}$$

With the definition (3.43) and according to (3.59) and (3.63) we deduce that

$$d_F(\mathscr{A}_\varepsilon) \leqq 2\left\{1+2K(\lambda_1|\Omega|)^{1/2}\left(G+\frac{\sqrt{2}}{16}\varepsilon G^2\right)\right\} \tag{3.64}$$

and since $\lambda_1|\Omega|$ is a nondimensional constant depending only on the shape of Ω, Theorem 3.2 follows in the case of homogeneous Dirichlet boundary conditions. As we indicated above, the computations are almost identical in the space periodic case and we omit them.

REFERENCES

[1] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I *and* II, Comm. Pure Appl. Math., 12 (1959), pp. 623-727; 17 (1964), pp. 35-92.

[2] M. BERCOVIER, *Perturbation of mixed variational problems. Applications to mixed finite element methods*, RAIRO Anal. Numér., 12 (1978).

[3] M. BERCOVIER AND M. S. ENGELMAN, *A finite element for numerical solution of viscous incompressible flow*, J. Comp. Phys., 20 (1986), pp. 181-201.

[4] P. CONSTANTIN, C. FOIAS AND R. TEMAM, *Attractors representing turbulent flows*, Mem. Amer. Math. Soc., 53 (1985).

[5] R. COURANT, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc. 49 (1943), pp. 1-23.

[6] C. FOIAS AND R. TEMAM, *Some analytic and geometric properties of the solutions of the Navier-Stokes equations*, J. Math. Pure Appl., 58 (1979), pp. 339-368.

[7] ———, unpublished.

[8] C. FOIAS, O. MANLEY AND R. TEMAM, *Attractors for the Bénard problem: existence and physical bounds on their fractal dimension*, Nonlinear Anal., 1988, to appear.

[9] E. LIEB AND W. THIRRING, *Inequalities for the moments of the eigenvalues of the Schrödinger equation and their relation to Sobolev inequalities*, in Studies in Mathematical Physics: Essays in Honor of Valentine Bargman, E. Lieb, B. Simon and A. S. Wightman, eds., Princeton Univ. Press, Princeton, NJ, 1976.

[10] B. MANDELBROT, *The Fractal Geometry of Nature*, W. H. Freeman, New York, 1983.

[11] J. T. ODEN AND N. KIKUCHI. *Penalty methods for constrained problems in elasticity*, Internat. J. Numer. Methods Engrg., 18 (1982), pp. 701-725.

[12] J. T. ODEN AND O. P. JACQUOTTE, *Stability of some mixed finite element methods for Stokesian flows*, Comput. Methods Appl. Mech. Engrg., 43 (1984), pp. 231-248.

[13] R. TEMAM, *Sur l'approximation des solutions des équations de Navier Stokes*, C.R. Acad. Sci. Paris Sér. A, 262 (1966), pp. 219-221.

[14] ———, *Une méthode d'approximation de la solution des équations de Navier-Stokes*, Bull. Soc. Math. France, 98 (1968), pp. 115-152.

[15] ———, *Navier Stokes Equations*, 3rd revised ed., North-Holland, Amsterdam-New York, 1984.

[16] ———, *Navier-Stokes Equations and Nonlinear Functional Analysis*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[17] ———, *Infinite dimensional dynamical systems in fluid mechanics*, in Nonlinear Functional Analysis and Applications, F. Browder, ed., Proc. Symposia in Pure Mathematics, Vol. 45, Part 2, 1982, pp. 431-445.

[18] ———, *Attractors for Navier Stokes equations*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. VII, H. Brezis and J. L. Lions, eds., Pitman, 1985, pp. 272-292.

[19] ———, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, Berlin, 1988, to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

002

ON THE MOTION OF VISCOUS FLUIDS IN THE PRESENCE OF DIFFUSION*

PAOLO SECCHI†

Abstract. We consider the motion of a viscous incompressible fluid consisting of two components with a diffusion effect obeying Fick's law. We prove: (i) the existence for two-dimensional flows of a (unique) global solution if the diffusion coefficient λ is small; (ii) the convergence (as $\lambda \to 0$) for two- and three-dimensional motions towards the corresponding solutions of the Navier-Stokes system for nonhomogeneous fluids.

Key words. Navier-Stokes equations, nonlinear parabolic equations

AMS(MOS) subject classifications. 35Q10, 76R99

1. Statement of the problem and main results. Let Ω be a bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$. Consider the motion in Ω of a viscous fluid consisting of two components, say, satured salt water and water. Let ρ_1, ρ_2 be the characteristic densities of the two components, $v^{(1)}(t,x)$ and $v^{(2)}(t,x)$ their velocities and $c(t,x)$, $d(t,x)$ the mass and volume concentration of the first fluid. The mean density of the mixture is $\rho(t,x) \equiv d\rho_1 + (1-d)\rho_2$ and the mean-volume and mean-mass velocities are $v(t,x) \equiv dv^{(1)} + (1-d)v^{(2)}$, $w(t,x) \equiv cv^{(1)} + (1-c)v^{(2)}$. Then the equations of motions are (see for instance Frank-Kamenestskii [3], Ignat'ev and Kuznetsov [4]):

$$\rho[\dot{w} + (w \cdot \nabla)w - b] - \mu \Delta w - (\mu + \mu')\nabla \operatorname{div} w = -\nabla \pi \quad \text{in } Q_T \equiv]0, T[\times \Omega,$$

$$\operatorname{div} v = 0 \quad \text{in } Q_T,$$

$$\dot{\rho} + \operatorname{div}(\rho w) = 0 \quad \text{in } Q_T,$$

where $\pi = \pi(t,x)$ is the (unknown) pressure, $b = b(t,x)$ is the external force field, μ and μ' ($\mu > 0, 3\mu' + 2\mu \geqq 0$) are the viscosity coefficients which are assumed to be constant. If the diffusion process obeys Fick's law

$$w = v - \lambda \frac{\nabla \rho}{\rho}$$

($\lambda > 0$ is the constant diffusion coefficient), we get

$$\begin{aligned} &\rho[\dot{v} + (v \cdot \nabla)v - b] - \mu \Delta v - \lambda[(v \cdot \nabla)\nabla\rho + (\nabla\rho \cdot \nabla)v] \\ &\quad + \nabla P + \frac{\lambda^2}{\rho}\left[(\nabla\rho \cdot \nabla)\nabla\rho - \frac{1}{\rho}(\nabla\rho \cdot \nabla\rho)\nabla\rho + \Delta\rho\nabla\rho\right] = 0 \quad \text{in } Q_T, \\ &\dot{\rho} + v \cdot \nabla\rho - \lambda\Delta\rho = 0 \quad \text{in } Q_T, \\ &\operatorname{div} v = 0 \quad \text{in } Q_T, \end{aligned} \tag{1.1}$$

where $P = \pi + \lambda v \cdot \nabla\rho - \lambda^2 \Delta\rho + \lambda(2\mu + \mu')\Delta \log \rho$ is the modified pressure. In addition, consider also the following initial and boundary conditions ($n = n(x)$ is the unit outward

* Received by the editors August 7, 1984; accepted for publication February 9, 1987.
† Università degli Studi di Trento, Dipartimento di Matematica, 38050 Povo (Trento), Italy.

normal to $\partial\Omega$):

$$
\begin{aligned}
&v = 0 && \text{on } \Sigma_T \equiv]0, T[\times \partial\Omega,\\
&\partial\rho/\partial n = 0 && \text{on } \Sigma_T,\\
&v|_{t=0} = v_0(x) && \text{in } \Omega,\\
&\rho|_{t=0} = \rho_0(x) && \text{in } \Omega.
\end{aligned} \tag{1.2}
$$

The initial density $\rho_0(x)$ is assumed to be a positive bounded function: $0 < m \leqq \rho_0(x) \leqq M$. In [6], [7], Kazhikhov and Smagulov study problem (1.1), (1.2) in the simplified case obtained from (1.1) by dropping the λ^2-terms. They prove the existence of a local (in time) solution (global in the bidimensional case) under the assumption

$$\lambda < \frac{2\mu}{\operatorname{osc} \rho_0}.$$

In [14], Smagulov and Utegenov study the asymptotic behavior as $t \to \infty$ of the solution of the simplified model; they also consider the problem of the behavior of the solution as $\lambda \to 0$. Local existence for the complete system (i.e., with the λ^2-terms), without any condition on λ, is proved by Secchi in [12] in the case $\Omega = \mathbb{R}^3$. More recently, Beirão da Veiga [2] has studied the complete problem (1.1), (1.2) in a bounded domain of $\mathbb{R}^3$. He has proved (with no assumption on λ) the existence of a (unique) local solution for every initial datum and external force field and the existence of a (unique) global solution if, as usual, the data are sufficiently small. In the present paper we consider the full system (1.1) and prove the existence of a (unique) global solution for the two-dimensional problem if λ/μ is sufficiently small. We need to introduce this condition in order to get the estimate of energy, necessary in our approach to obtain the global existence. Moreover we consider the behavior of the solution as $\lambda \to 0$. We prove that there exists a subsequence of solutions of (1.1), (1.2) converging towards a solution of the corresponding Navier-Stokes equations for nonhomogeneous incompressible fluids:

$$
\begin{aligned}
&\rho[\dot{v} + (v \cdot \nabla)v - b] - \mu\Delta v + \nabla P = 0 \quad \text{in } Q_T,\\
&\dot{\rho} + v \cdot \nabla\rho = 0 \quad \text{in } Q_T,\\
&\operatorname{div} v = 0 \quad \text{in } Q_T,\\
&v = 0 \quad \text{on } \Sigma_T,\\
&v|_{t=0} = v_0(x) \quad \text{in } \Omega,\\
&\rho|_{t=0} = \rho_0(x) \quad \text{in } \Omega.
\end{aligned} \tag{1.3}
$$

If the domain is two-dimensional the convergence is on every finite interval $[0, T]$. In the three-dimensional case, where the solution exists only in the small, the convergence is on a short time interval independent of λ. Concerning the Navier-Stokes equations for nonhomogeneous fluids see Antoncev and Kazhikhov [1], Kazhikhov [5], Ladyzhenskaja and Solonnikov [9]; see also Lions [10] and Simon [13]. We find the solution to the Navier-Stokes system in the same class of solutions of Kazhikhov.

Let us denote by $\| \ \|$ and $(\ , \ )$ the norm and the scalar product in $L^2(\Omega)$, by $| \ |_p$ the norm in $L^p(\Omega)$, $2 < p \leqq \infty$. We use the same notation for scalar and vector spaces.

Let $H^k(\Omega)$ be the usual Sobolev space and $\| \ \|_k$ its norm. Let $H_0^1(\Omega)$ be the closure of $C_0^\infty(\Omega)$ in $H^1(\Omega)$. We introduce the following functional spaces:

$$H_N^k \equiv \left\{ \rho \in H^k(\Omega): \frac{\partial \rho}{\partial n} = 0 \text{ on } \partial\Omega, \int_\Omega \rho(x)dx = \int_\Omega \rho_0(x)dx \right\}, \qquad k \geqq 2,$$

$$V \equiv \{v \in C_0^\infty(\Omega): \operatorname{div} v = 0 \text{ in } \Omega\},$$

$$H \equiv \{v \in L^2(\Omega): \operatorname{div} v = 0 \text{ in } \Omega, \ v \cdot n = 0 \text{ on } \partial\Omega\},$$

$$V \equiv \{v \in H_0^1(\Omega): \operatorname{div} v = 0 \text{ in } \Omega\}$$

(see for instance [8], [11], [15] for their properties). H and V are the closures of V in $L^2(\Omega)$ and $H_0^1(\Omega)$, respectively. Moreover $L^2(\Omega) = H \oplus G$, where $G \equiv \{\nabla p: p \in H^1(\Omega)\}$. Denoting by P_H the orthogonal projection of $L^2(\Omega)$ onto H, we define the operator $A \equiv -P_H \Delta$ on $D(A) \equiv H^2(\Omega) \cap V$. We have

$$(Au, v) = \sum_{i,j} (D_i u_j, D_i v_j) \quad \forall u \in D(A), \quad v \in V$$

where D_i means $\partial/\partial x_i$. The norms $\|\rho\|_2$, $\|\Delta\rho\|$ are equivalent in H_N^2, $\|\rho\|_3$, $\|\nabla\Delta\rho\|$ are equivalent in H_N^3 and $\|v\|_2$, $\|Av\|$ are equivalent in $D(A)$. We define $\|v\|_V^2 \equiv \sum_{i,j} (D_i v_j, D_i v_j)$; the norms $\|v\|_1$, $\|v\|_V$ are equivalent in V. If X is a Banach space, $L^2(0, T; X)$ will be the Banach space of X-valued measurable functions in $L^2(0, T)$, and $C(0, T; X)$ will be the Banach space of X-valued continuous functions on $[0, T]$. We set $L^2(Q_T) \equiv L^2(0, T; L^2(\Omega))$, $H^1(0, T; L^2(\Omega)) \equiv \{v \in L^2(Q_T)$ with $\dot{v} \in L^2(Q_T)\}$. In the sequel C will denote different constants depending at most on Ω and m, M, μ. Other constants will be indicated by C_0, C_1, C_2, $\cdots$. We prove the following results.

THEOREM A. *Let Ω be an open bounded set in $\mathbb{R}^2$ with boundary $\partial\Omega$ of class C^3. Suppose $v_0 \in V$, $\rho_0 \in H_N^2$ such that $0 < m \leqq \rho_0(x) \leqq M$ in Ω, $b \in L^2(Q_T)$. Then there exists $\lambda_0 > 0$ depending on Ω, m, M such that, if $\lambda/\mu < \lambda_0$, problem (1.1), (1.2) is uniquely solvable in Q_T. Moreover $v \in L^2(0, T; H^2(\Omega)) \cap C(0, T; V)$, $\dot{v} \in L^2(0, T; H)$, $\rho \in L^2(0, T; H_N^3) \cap C(0, T; H_N^2)$, $\dot{\rho} \in L^2(0, T; H^1(\Omega))$ and $m \leqq \rho(t, x) \leqq M$, $\nabla P \in L^2(Q_T)$.*

DEFINITION. We shall say that $(v, \rho, \nabla P)$ is a *generalized solution* of (1.3) if:

(i) $v \in L^2(0, T; H^2(\Omega)) \cap H^1(0, T; L^2(\Omega))$, $\rho \in L^\infty(Q_T)$, $\nabla P \in L^2(Q_T)$;

(ii) $(1.3)_1$, $(1.3)_3$–$(1.3)_5$ are satisfied in the usual strong sense and $(1.3)_2$, $(1.3)_6$ are satisfied in the following weak sense:

$$\int_0^T (\rho, \dot{\varphi} + v \cdot \nabla\varphi) \, dt + (\rho_0, \varphi(0, \cdot)) = 0 \tag{1.4}$$

for any $\varphi \in H^1(Q_T)$ such that $\varphi(T, x) = 0$.

THEOREM B. *Let Ω be an open bounded set in $\mathbb{R}^n$, $n = 2$ or 3, with $\partial\Omega$ of class C^3. Assume $v_0 \in V$, $\rho_0 \in H_N^2$ such that $0 < m \leqq \rho_0(x) \leqq M$, in Ω, $b \in L^2(Q_{T_0})$. Then there exist $T \in (0, T_0]$ (independent of λ) and a subsequence of $\{(v^\lambda, \rho^\lambda, \nabla P^\lambda)\}$ converging on $[0, T]$ in the topology indicated in (3.12), (3.13) to a generalized solution $(v, \rho, \nabla P)$ of (1.3). The function ρ satisfies $m \leqq \rho(t, x) \leqq M$ a.e. in Q_T. If $n = 2$, $T = T_0$.*

Remark. It is not known if generalized solutions of (1.3) are unique. In [9], Ladyzhenskaja and Solonnikov show the existence of a solution of (1.3) $v \in L^q(0, T; W_q^2(\Omega)) \cap W_q^1(0, T; L^q(\Omega))$, $\nabla P \in L^q(Q_T)$, $q > n$ $(n = 2, 3)$, $\rho \in C^1(\bar{Q}_T)$, provided that $v_0 \in W_q^{2-2/q}(\Omega)$, $\rho_0 \in C^1(\bar{\Omega})$, $b \in L^q(Q_T)$. Here T is sufficiently small in $n = 3$, arbitrary if $n = 2$. In this class of functions the uniqueness theorem holds. It is also

possible to prove that these smoother solutions (if the data are regular enough so that they exist) necessarily coincide with the generalized solutions. Hence, if the data are smooth enough, the uniqueness theorem holds also for generalized solutions. This implies the convergence of the whole sequence $\{(v^\lambda, \rho^\lambda, \nabla P^\lambda)\}$ to the solution $(v, \rho, \nabla P)$.

2. Proof of Theorem A. Let $(v, \rho, \nabla P)$ be a solution to problem (1.1), (1.2). We want to find some global a priori estimate for it. In view of the study of the next section (the convergence as $\lambda \to 0$), we shall explicitly point out every dependence on λ, also if, for the aim of this section, λ is considered fixed. We start with the estimates for the density. From $(1.1)_2$ and $(1.1)_3$, $(1.2)_1$, $(1.2)_2$ one has $d/dt \int_\Omega \rho\, dx = 0$; hence $\int_\Omega \rho\, dx = \int_\Omega \rho_0\, dx$. From the diffusion equation $(1.1)_2$, we have by the maximum principle

$$0 < m \leqq \rho(t, x) \leqq M, \qquad (t, x) \in Q_T, \tag{2.1}$$

for any $\lambda > 0$. Multiply $(1.1)_2$ by $-\lambda \Delta \rho$ and integrate over Ω. Then, by integrating by parts and using $(1.1)_3$, $(1.2)_1$, $(1.2)_2$, we obtain

$$\frac{\lambda}{2} \frac{d}{dt} \|\nabla \rho\|^2 + \lambda^2 \|\Delta \rho\|^2 + \lambda((\nabla \rho \cdot \nabla) v, \nabla \rho) = 0. \tag{2.2}$$

Using (2.1) and the interpolation inequality

$$|\nabla \rho|_4^2 \leqq C_0 \|\Delta \rho\| \, |\rho|_\infty, \tag{2.3}$$

where C_0 is a positive constant depending on Ω, we easily obtain

$$\frac{\lambda}{2} \frac{d}{dt} \|\nabla \rho\|^2 + \lambda^2 \|\Delta \rho\|^2 \leqq \varepsilon_1 \lambda^2 \|\Delta \rho\|^2 + \frac{C_0^2 M^2}{4 \varepsilon_1} \|v\|_V^2 \tag{2.4}$$

where ε_1 is a small positive parameter. From $(1.1)_1$ and $(1.1)_2$ we have, after suitable integrations by parts,

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} (\rho v, v) &= (\rho b, v) - \mu \|v\|_V^2 - \lambda((v \cdot \nabla) v, \nabla \rho) \\ &\quad - \left(\frac{\lambda^2}{\rho} \left[(\nabla \rho \cdot \nabla) \nabla \rho - \frac{1}{\rho} (\nabla \rho \cdot \nabla \rho) \nabla \rho + \Delta \rho \nabla \rho \right], v \right). \end{aligned} \tag{2.5}$$

On the other hand, by integrating by parts, we have

$$-\left(\frac{\lambda^2}{\rho} (\nabla \rho \cdot \nabla) \nabla \rho, v \right) = \left(\frac{\lambda^2}{\rho} \Delta \rho \nabla \rho, v \right) + \left(\frac{\lambda^2}{\rho} (\nabla \rho \cdot \nabla) v, \nabla \rho \right) - \left(\frac{\lambda^2}{\rho^2} |\nabla \rho|^2 \nabla \rho, v \right).$$

Then (2.5) becomes

$$\frac{1}{2} \frac{d}{dt} (\rho v, v) + \mu \|v\|_V^2 = (\rho b, v) - \lambda((v \cdot \nabla) v, \nabla \rho) + \left(\frac{\lambda^2}{\rho} (\nabla \rho \cdot \nabla) v, \nabla \rho \right). \tag{2.6}$$

Set $c_0 = (M + m)/2$ so that $|\rho(t, x) - c_0| \leqq (M - m)/2$. We have

$$\begin{aligned} \lambda |((v \cdot \nabla) v, \nabla \rho)| &= \lambda |((v \cdot \nabla) v, \nabla (\rho - c_0))| \\ &= \lambda \left| \sum_{i,j} (D_i v_j, D_i v_j (\rho - c_0)) \right| \leqq \lambda \frac{M - m}{2} \|v\|_V^2. \end{aligned}$$

Moreover, by using (2.1), (2.3) and introducing a small parameter ε_2,

$$\left|\left(\frac{\lambda^2}{\rho}(\nabla\rho\cdot\nabla)v, \nabla\rho\right)\right| \leqq \varepsilon_2\lambda^2\|\Delta\rho\|^2+\frac{C_0^2M^2\lambda^2}{4\varepsilon_2 m^2}\|v\|_V^2.$$

Then (2.6) gives

$$\frac{1}{2}\frac{d}{dt}(\rho v, v)+\mu\|v\|_V^2 \leqq (\rho b, v)+\left(\frac{M-m}{2}\lambda+\frac{C_0^2M^2}{4\varepsilon_2 m^2}\lambda^2\right)\|v\|_V^2+\varepsilon_2\lambda^2\|\Delta\rho\|^2. \tag{2.7}$$

Now we have to balance (2.4) and (2.7) in a suitable way. We can proceed in the following way: We add (2.7) to (2.4) multiplied by $\alpha=\mu/2C_0^2M^2$ and set $\varepsilon_1=1/4$, $\varepsilon_2=\alpha/4$. Then there exists $\lambda_0>0$ sufficiently small depending on m, M, C_0 such that if $\lambda/\mu<\lambda_0$ we have

$$\frac{1}{2}\frac{d}{dt}[(\rho v, v)+\alpha\lambda\|\nabla\rho\|^2]+\frac{\alpha}{2}\lambda^2\|\Delta\rho\|^2+\frac{\mu}{4}\|v\|_V^2 \leqq (\rho b, v).$$

By Gronwall's lemma and (2.1), after some calculation, we get for any $\lambda<\lambda_0\mu$

$$\begin{aligned}\|v\|_{C(0,T;H)}+\lambda^{1/2}\|\nabla\rho\|_{C(0,T;L^2(\Omega))}+\lambda\|\Delta\rho\|_{L^2(Q_T)}+\|v\|_{L^2(0,T;V)}\\ \leqq C[\|v_0\|+\|\nabla\rho_0\|+T^{1/2}\|b\|_{L^2(Q_T)}].\end{aligned} \tag{2.8}$$

Now we shall prove a priori estimates for higher norms. From $(1.1)_2$ and $(1.2)_2$, we have

$$\frac{1}{2}\frac{d}{dt}\|\Delta\rho\|^2=\int_{\partial\Omega}\Delta\rho\frac{\partial\dot\rho}{\partial n}\,d\sigma-(\nabla\Delta\rho, \nabla\dot\rho)=(\nabla\Delta\rho, \nabla(v\cdot\nabla\rho-\lambda\Delta\rho)).$$

Integrating by parts and using $(1.1)_3$ and $(1.2)_1$, we obtain

$$(\nabla\Delta\rho, \nabla(v\cdot\nabla\rho))=((\nabla\Delta\rho\cdot\nabla)v, \nabla\rho)-\sum_{i,j,k}(D_jv_i, D_kD_i\rho D_kD_j\rho).$$

Then we have

$$\frac{\lambda^2}{2}\frac{d}{dt}\|\Delta\rho\|^2+\lambda^3\|\nabla\Delta\rho\|^2=\lambda^2((\nabla\Delta\rho\cdot\nabla)v, \nabla\rho)-\lambda^2\sum_{i,j,k}(D_jv_i, D_kD_i\rho D_kD_j\rho). \tag{2.9}$$

Using (2.3) and the interpolation inequality

$$|\nabla v|_4^2 \leqq C\|v\|_2\|\nabla v\| \leqq C\|Av\|\,\|v\|_V$$

we obtain

$$\lambda^2|((\nabla\Delta\rho\cdot\nabla)v, \nabla\rho)| \leqq \frac{\varepsilon_0}{4}\|Av\|^2+\frac{\varepsilon}{4}\lambda^3\|\nabla\Delta\rho\|^2+\frac{C}{\varepsilon_0\varepsilon^2}\|v\|_V^4+\frac{C}{\varepsilon_0\varepsilon^2}\lambda^4\|\Delta\rho\|^4$$

where ε_0 and ε are small positive parameters (say less than 1). On the other hand, by using the inequalities

$$\begin{aligned}|\nabla v|_3 &\leqq C\|v\|_2^{1/3}\|\nabla v\|^{2/3} \leqq C\|Av\|^{1/3}\|v\|_V^{2/3},\\ |\Delta\rho|_3 &\leqq C\|\nabla\Delta\rho\|^{2/3}|\rho|_\infty^{1/3},\end{aligned}$$

we have (since $ab \leqq \varepsilon a^{3/2}+(C/\varepsilon^2)b^3$)

$$\lambda^2\left|\sum_{i,j,k}(D_jv_i, D_kD_i\rho D_kD_j\rho)\right| \leqq \frac{\varepsilon_0}{4}\|Av\|^2+\frac{\varepsilon}{4}\lambda^3\|\nabla\Delta\rho\|^2+\frac{C}{\varepsilon_0\varepsilon^4}\|v\|_V^4. \tag{2.10}$$

We thus obtain

$$\frac{\lambda^2}{2}\frac{d}{dt}\|\Delta\rho\|^2+\lambda^3\|\nabla\Delta\rho\|^2 \leqq \frac{\varepsilon_0}{2}\|Av\|^2+\frac{\varepsilon}{2}\lambda^3\|\nabla\Delta\rho\|^2+\frac{C}{\varepsilon_0\varepsilon^4}\|v\|_V^4+\frac{C}{\varepsilon_0\varepsilon^2}\lambda^4\|\Delta\rho\|^4. \tag{2.11}$$

Now we turn to the estimate for the velocity. We obtain it by following the method used in [2]. If we take the projection P_H of $(1.1)_1$, we obtain

$$P_H(\rho\dot{v}) - \mu Av = F \tag{2.12}$$

where for the sake of brevity

$$F \equiv P_H\Big\{-\rho(v\cdot\nabla)v + \rho b + \lambda[(v\cdot\nabla)\nabla\rho + (\nabla\rho\cdot\nabla)v] - \frac{\lambda^2}{\rho}\Big[(\nabla\rho\cdot\nabla)\nabla\rho - \frac{1}{\rho}(\nabla\rho\cdot\nabla\rho)\nabla\rho + \Delta\rho\nabla\rho\Big]\Big\}.$$

In H take the inner product of (2.12) with $\dot{v} + (m\mu/4M^2)Av$. Since

$$(\dot{v}, Av) = \frac{1}{2}\frac{d}{dt}\|v\|_V^2,$$

we obtain

$$m\|\dot{v}\|^2 + \frac{\mu}{2}\frac{d}{dt}\|v\|_V^2 + \frac{m\mu^2}{4M^2}\|Av\|^2 \leqq \|F\|\,\|\dot{v}\| + \frac{m\mu}{4M^2}\|F\|\,\|Av\| + \frac{m\mu}{4M}\|\dot{v}\|\,\|Av\|,$$

from which we easily obtain

$$\mu\frac{d}{dt}\|v\|_V^2 + m\|\dot{v}\|^2 + \frac{m\mu^2}{4M^2}\|Av\|^2 \leqq \left(\frac{2}{m} + \frac{m}{2M^2}\right)\|F\|^2. \tag{2.13}$$

We have only to estimate the norm of F in $L^2(\Omega)$. By using Hölder's inequality and suitable interpolation inequalities, we obtain

$$\begin{aligned}\|F\|^2 \leqq{}& C\|v\|\,\|v\|_V^2\|Av\| + C\lambda^2\|v\|^{2/3}\|v\|_V^{4/3}\|\nabla\Delta\rho\|^{4/3} + C\lambda^2\|\Delta\rho\|\,\|Av\|\,\|v\|_V \\ &+ C\lambda^4\|\Delta\rho\|\,\|\nabla\Delta\rho\|^{3/2} + C\lambda^4\|\Delta\rho\|^4 + C\|b\|^2.\end{aligned}$$

After some calculation we then get from (2.13) the following estimate:

$$\begin{aligned}&\frac{d}{dt}\|v\|_V^2 + \frac{m}{\mu}\|\dot{v}\|^2 + \frac{m\mu}{4M^2}\|Av\|^2 \leqq \varepsilon_0\|Av\|^2 + \varepsilon\lambda^3\|\nabla\Delta\rho\|^2 \\ &\quad + C\left(\frac{1}{\varepsilon_0} + \frac{1}{\varepsilon^2}\right)(1+\|v\|^2)\|v\|_V^4 + C\left(\frac{1}{\varepsilon_0} + \frac{1}{\varepsilon^3}\right)\lambda^4\|\Delta\rho\|^4 + C\|b\|^2.\end{aligned} \tag{2.14}$$

Finally, adding (2.11) to (2.14) divided by 2, we have

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}[\|v\|_V^2 + \lambda^2\|\Delta\rho\|^2] + \frac{m}{2\mu}\|\dot{v}\|^2 + \frac{m\mu}{8M^2}\|Av\|^2 + \lambda^3\|\nabla\Delta\rho\|^2 \\ &\quad \leqq \varepsilon_0\|Av\|^2 + \varepsilon\lambda^3\|\nabla\Delta\rho\|^2 + \frac{C}{\varepsilon_0\varepsilon^4}(1+\|v\|^2)\|v\|_V^4 + \frac{C}{\varepsilon_0\varepsilon^3}\lambda^4\|\Delta\rho\|^4 + C\|b\|^2.\end{aligned}$$

By taking ε_0 and ε sufficiently small, we thus obtain

$$\begin{aligned}&\frac{d}{dt}[\|v\|_V^2 + \lambda^2\|\Delta\rho\|^2] + C[\|\dot{v}\|^2 + \|Av\|^2 + \lambda^3\|\nabla\Delta\rho\|^2] \\ &\quad \leqq C[(1+\|v\|^2)\|v\|_V^2 + \lambda^2\|\Delta\rho\|^2][\|v\|_V^2 + \lambda^2\|\Delta\rho\|^2] + C\|b\|^2.\end{aligned} \tag{2.15}$$

By (2.8) the function $(1+\|v\|^2)\|v\|_V^2 + \lambda^2\|\Delta\rho\|^2$ belongs to $L^1(0, T)$. This gives, for any

$\lambda < \lambda_0 \mu$, the following estimate:

$$\begin{aligned} &\|v\|_{C(0,T;V)} + \lambda \|\Delta\rho\|_{C(0,T;L^2(\Omega))} + \|\dot{v}\|_{L^2(0,T;H)} + \|Av\|_{L^2(0,T;H)} \\ &\quad + \lambda^{3/2} \|\nabla \Delta\rho\|_{L^2(Q_T)} \leqq K_1(\|v_0\|_V + \|\nabla\rho_0\| \\ &\quad + T^{1/2}\|b\|_{L^2(Q_T)}) \cdot [\|v_0\|_V + \|\Delta\rho_0\| + \|b\|_{L^2(Q_T)}] \end{aligned} \tag{2.16}$$

where K_1 is an increasing function of its argument depending also on m, M, μ, Ω. Finally, directly from $(1.1)_1$, we can obtain $\nabla P \in L^2(Q_T)$, and from $(1.1)_2$ we can obtain $\dot{\rho} \in L^2(0, T; H^1(\Omega))$. By using these estimates, it is possible to prove the existence in the large of a solution of (1.1), (1.2). For example, one can follow (with obvious modifications) the continuity method employed in [2]. Also the uniqueness can be proved as in [2].

Remark. If $b \in L^1(0, +\infty; L^2(\Omega)) \cap L^2(0, +\infty; L^2(\Omega))$ it is possible to get the existence of the solution on the whole interval $[0, +\infty)$. The assumption $b \in L^1(0, +\infty; L^2(\Omega))$ is necessary to obtain the estimate (2.8) for $T = +\infty$, with on the right $C[\|v_0\| + \|\nabla\rho_0\| + \|b\|_{L^1(0,+\infty;L^2(\Omega))}]$. Analogously, also (2.16) will be true for $T = +\infty$, with on the right

$$K_1(\|v_0\|_V + \|\nabla\rho_0\| + \|b\|_{L^1(0,+\infty;L^2(\Omega))}) \cdot [\|v_0\|_V + \|\Delta\rho_0\| + \|b\|_{L^2(0,+\infty;L^2(\Omega))}].$$

These estimates permit to extend the solution from every bounded interval $[0, T]$ to the whole interval $[0, +\infty)$.

3. Proof of Theorem B. Let us now consider the problem of the behavior of the solution to problem (1.1), (1.2) as $\lambda \to 0$. For the passage to the limit we need suitable estimates not depending on λ. We already proved these estimates in § 2 for the bidimensional case. Now, we shall prove them for the three-dimensional case. Let us denote by $(v^\lambda, \rho^\lambda, \nabla P^\lambda)$ the solution of (1.1), (1.2) (existence and uniqueness of this solution is proved in [2]). This solution exists in a short interval which is, a priori, dependent on λ. We shall see that one can find an interval not depending on λ; the convergence will be proved in this time interval. Assume $0 < \lambda \leqq \lambda_1$ with $\lambda_1 > 0$ arbitrary. We proceed as in § 2. From the diffusion equation we obtain

$$0 < m \leqq \rho^\lambda(t, x) \leqq M, \qquad (t, x) \in Q_T, \tag{3.1}$$

for any $\lambda > 0$. As in the bidimensional case we obtain (the analogue of (2.4)):

$$\frac{\lambda}{2}\frac{d}{dt}\|\nabla\rho^\lambda\|^2 + \lambda^2\|\Delta\rho^\lambda\|^2 \leqq \frac{\lambda^2}{2}\|\Delta\rho^\lambda\|^2 + C\|v^\lambda\|_V^2 \tag{3.2}$$

(observe that (2.3) is valid also for three-dimensional domains). From (2.9), by using (3.1) and the following interpolation inequalities

$$|\nabla v|_3 \leqq C\|\Delta v\|^{1/2}\|\nabla v\|^{1/2} \leqq C\|Av\|^{1/2}\|v\|_V^{1/2}, \tag{3.3}$$

$$|\nabla\rho|_6 \leqq C\|\nabla\Delta\rho\|^{1/3}|\rho|_\infty^{2/3}, \tag{3.4}$$

$$|D^2\rho|_3 \leqq C\|\nabla\Delta\rho\|^{2/3}|\rho|_\infty^{1/3}, \tag{3.5}$$

we get

$$\begin{aligned} \frac{\lambda^2}{2}\frac{d}{dt}\|\Delta\rho^\lambda\|^2 + \lambda^3\|\nabla\Delta\rho^\lambda\|^2 &\leqq \lambda^2\|\nabla\Delta\rho^\lambda\|\,|\nabla v^\lambda|_3|\nabla\rho^\lambda|_6 + C\lambda^2|\nabla v^\lambda|_3|D^2\rho^\lambda|_3^2 \\ &\leqq C\lambda^2\|\Delta v^\lambda\|^{1/2}\|\nabla v^\lambda\|^{1/2}\|\nabla\Delta\rho^\lambda\|^{4/3}|\rho^\lambda|_\infty^{2/3} \\ &\leqq \frac{\lambda^3}{4}\|\nabla\Delta\rho^\lambda\|^2 + \varepsilon\|Av^\lambda\|^2 + \frac{C}{\varepsilon^3}\|v^\lambda\|_V^6 \end{aligned} \tag{3.6}$$

where ε is a small parameter. Finally, consider the estimate (2.13), where instead of v and F we write v^λ and F^λ. By using (3.1), (3.3)-(3.5) and the Sobolev embedding theorem $H^1 \subset L^6$, we have

$$\|F^\lambda\|^2 \leqq C\|v^\lambda\|_V^3\|Av^\lambda\| + C\lambda^2\|v^\lambda\|_V^2\|\nabla\Delta\rho^\lambda\|^{4/3} + C\lambda^2\|Av^\lambda\|\,\|v^\lambda\|_V\|\nabla\Delta\rho^\lambda\|^{2/3} + C\lambda^4\|\nabla\Delta\rho^\lambda\|^{4/3}\|\Delta\rho^\lambda\|^2 + C\|b\|^2.$$

From this estimate we obtain (for $\varepsilon > 0$ small), after some calculation,

$$\begin{aligned}&\frac{d}{dt}\|v^\lambda\|_V^2 + C\|\dot v^\lambda\|^2 + C\|Av^\lambda\|^2 \\ &\qquad \leqq \varepsilon\|Av^\lambda\|^2 + \frac{\lambda^3}{2}\|\nabla\Delta\rho^\lambda\|^2 + \frac{C}{\varepsilon^3}\|v^\lambda\|_V^6 + \frac{C}{\varepsilon^3}\lambda^6\|\Delta\rho^\lambda\|^6 + C\|b\|^2\end{aligned} \tag{3.7}$$

where the constants C in the third and in the fourth term of the right side depend also on λ_1. Thus, from (3.6) and (3.7), it is possible to obtain the following estimate:

$$\begin{aligned}&\frac{d}{dt}[\|v^\lambda\|_V^2 + \lambda^2\|\Delta\rho^\lambda\|^2] + C\|\dot v^\lambda\|^2 + (C - 3\varepsilon)\|Av^\lambda\|^2 + \lambda^3\|\nabla\Delta\rho^\lambda\|^2 \\ &\qquad \leqq \frac{C}{\varepsilon^3}[\|v^\lambda\|_V^2 + \lambda^2\|\Delta\rho^\lambda\|^2]^3 + C\|b\|^2.\end{aligned} \tag{3.8}$$

Let ε be such that $C - 3\varepsilon > 0$. Then there exist $T' \in (0, T_0]$ and a constant C_1 depending only on m, M, μ, Ω, $\|v_0\|_V$, $\lambda_1\|\Delta\rho_0\|$ and $\|b\|_{L^2(Q_T)}$ such that

$$\begin{aligned}&\|v^\lambda\|_{C(0,T';V)} + \lambda\|\Delta\rho^\lambda\|_{C(0,T';L^2(\Omega))} + \lambda^{3/2}\|\nabla\Delta\rho^\lambda\|_{L^2(Q_{T'})} + \|\dot v^\lambda\|_{L^2(Q_{T'})} \\ &\qquad + \|Av^\lambda\|_{L^2(Q_{T'})} \leqq C_1,\end{aligned} \tag{3.9}$$

for any $0 < \lambda < \lambda_1$. From (3.2) we see that there exists a constant C_2, depending on the data of the problem as C_1, such that

$$\lambda^{1/2}\|\nabla\rho^\lambda\|_{C(0,T';L^2(\Omega))} \leqq C_2 \tag{3.10}$$

for any $0 < \lambda < \lambda_1$. Directly from $(1.1)_1$ we obtain

$$\|\nabla P^\lambda\|_{L^2(Q_{T'})} \leqq C_3 \tag{3.11}$$

where C_3 is a positive constant depending on the data as C_1.

The convergence as $\lambda \to 0$. From now on we treat both the two-dimensional and the three-dimensional cases together. Let T be arbitrary in the plane case and $T = T'$ in the spatial case. Since $\{v^\lambda\}$ is bounded in $L^2(0, T; H^2(\Omega)) \cap H^1(0, T; L^2(\Omega))$, it is bounded by interpolation in $C(0, T; H^1(\Omega)) \cap H^{1/2+\varepsilon}(0, T; H^{1-\varepsilon}(\Omega))$; hence from the Ascoli-Arzelà theorem it is compact in $C(0, T; H^{1-\varepsilon}(\Omega))$. Then there exists a subsequence, that we continue to denote by $\{v^\lambda\}$, and a function v such that

$$\begin{aligned}&v^\lambda \to v \quad \text{in } C(0, T; H^{1-\varepsilon}(\Omega)), \\ &v^\lambda \rightharpoonup v \quad \text{in } L^2(0, T; H^2(\Omega)) \text{ and in } H^1(0, T; L^2(\Omega)).\end{aligned} \tag{3.12}$$

From (2.1), (3.1) and the estimates on the pressures P^λ, we find two subsequences, again denoted by $\{\rho^\lambda\}$, $\{P^\lambda\}$, and two functions ρ, P such that

$$\begin{aligned}&\rho^\lambda \rightharpoonup^* \rho \quad \text{in } L^\infty(Q_T), \\ &\nabla P^\lambda \rightharpoonup \nabla P \quad \text{in } L^2(Q_T).\end{aligned} \tag{3.13}$$

First we pass to the limit in $(1.1)_2$, where instead of $(v, \rho, \nabla P)$ we consider $(v^\lambda, \rho^\lambda, \nabla P^\lambda)$. Take $\varphi \in H^1(Q_T)$ with $\varphi(T, x) = 0$. Then

$$0 = \int_0^T (\dot{\rho}^\lambda + v^\lambda \cdot \nabla \rho^\lambda - \lambda \Delta \rho^\lambda, \varphi)\, dt$$

$$= -\int_0^T (\rho^\lambda, \dot{\varphi} + v^\lambda \cdot \nabla \varphi)\, dt + \lambda \int_0^T (\nabla \rho^\lambda, \nabla \varphi)\, dt - (\rho_0, \varphi(0, \cdot));$$

using the strong convergence of v^λ and the weak-$*$ convergence of ρ^λ, estimates (2.8) and (3.10), one can pass to the limit as $\lambda \to 0$ in the right-hand side of this equation obtaining (1.4). Consider now $(1.1)_1$. By $(1.1)_2$ and some integrations by parts we obtain, for any $\phi \in C_0^\infty(Q_T)$,

$$\int_0^T (\rho^\lambda \dot{v}^\lambda + \rho^\lambda (v^\lambda \cdot \nabla) v^\lambda, \phi)\, dt = -\int_0^T (\rho^\lambda v^\lambda, \dot{\phi})\, dt - \int_0^T (\rho^\lambda (v^\lambda \cdot \nabla)\phi, v^\lambda)\, dt$$

(3.14)

$$+ \int_0^T ((\nabla \rho^\lambda \cdot \nabla) v^\lambda, \phi)\, dt + \lambda \int_0^T ((\nabla \rho^\lambda \cdot \nabla)\phi, v^\lambda)\, dt.$$

By (2.8), (2.16), (3.9) and (3.10) the last two terms can be estimated by

$$\lambda^{1/2} \lambda^{1/2} \|\nabla \rho^\lambda\|_{C(0,T;L^2(\Omega))} \|v^\lambda\|_{C(0,T;V)} \|\phi\|_{L^\infty(Q_T)} T$$

$$+ \lambda^{1/2} \lambda^{1/2} \|\nabla \rho^\lambda\|_{C(0,T;L^2(\Omega))} \|\nabla \phi\|_{L^\infty(Q_T)} \|v^\lambda\|_{C(0,T;L^2(\Omega))} T \leqq C_4 \lambda^{1/2},$$

going to zero as $\lambda \to 0$. By using the strong convergence v^λ and the weak-$*$ convergence of ρ^λ, we can pass to the limit in the other terms of (3.14) obtaining

$$\int_0^T (\rho^\lambda \dot{v}^\lambda + \rho^\lambda (v^\lambda \cdot \nabla) v^\lambda, \phi)\, dt \to -\int_0^T (\rho v, \dot{\phi})\, dt - \int_0^T (\rho (v \cdot \nabla)\phi, v)\, dt$$

for any $\phi \in C_0^\infty(Q_T)$. Consider now $\varphi = v \cdot \phi$; then $\varphi \in H^1(Q_T)$, $\varphi(T, x) = \varphi(0, x) = 0$. By (1.4) we have

$$-\int_0^T (\rho v, \dot{\phi})\, dt - \int_0^T (\rho (v \cdot \nabla)\phi, v)\, dt = \int_0^T (\rho \dot{v} + \rho (v \cdot \nabla) v, \phi)\, dt.$$

The convergence of $\int_0^T (\rho^\lambda b + \mu \Delta v^\lambda - \nabla P^\lambda, \phi)\, dt$ to the corresponding expression is direct. Consider the first term in λ in $(1.1)_1$. We have, by integration by parts and (3.10),

$$\left| \lambda \int_0^T ((v^\lambda \cdot \nabla) \nabla \rho^\lambda, \phi)\, dt \right| = \left| \lambda \int_0^T ((v^\lambda \cdot \nabla)\phi, \nabla \rho^\lambda)\, dt \right|$$

$$\leqq \lambda^{1/2} T \|\nabla \phi\|_{L^\infty(Q_T)} \|v\|_{L^\infty(0,T;L^2(\Omega))} \lambda^{1/2} \|\nabla \rho^\lambda\|_{L^\infty(0,T;L^2(\Omega))} \leqq C_4 \lambda^{1/2},$$

going to zero as $\lambda \to 0$. Concerning the second term in λ^2, we have, by using (3.4) and (3.9),

$$\left| \lambda \int_0^T \left(\frac{1}{(\rho^\lambda)^2} (\nabla \rho^\lambda \cdot \nabla \rho^\lambda) \nabla \rho^\lambda, \phi \right) dt \right| \geqq \lambda^{1/2} C \frac{M^2}{m^2} \|\phi\|_{L^2(Q_T)} \lambda^{3/2} \|\nabla \Delta \rho^\lambda\|_{L^2(Q_T)} \leqq C_5 \lambda^{1/2},$$

which goes to zero as $\lambda \to 0$. Also the other terms in λ, λ^2 go to zero as $\lambda \to 0$, as one can see with a direct calculation. Thus at the limit we obtain

$$\int_0^T (\rho \dot{v} + \rho (v \cdot \nabla) v - \rho b - \mu \Delta v + \nabla P, \phi)\, dt = 0$$

for any $\phi \in C_0^\infty(Q_T)$. This gives $(1.3)_1$. $(1.3)_3$ and the initial and boundary conditions $(1.3)_4$, $(1.3)_5$ are easily checked. The boundary condition $(1.2)_2$ on ρ^λ is lost.

REFERENCES

[1] S. N. ANTONCEV AND A. V. KAZHIKHOV, *Mathematical study of flows of non-homogeneous fluids*, Novosibirsk, Lecture at the University, 1973.

[2] H. BEIRÃO DA VEIGA, *Diffusion on viscous fluids. Existence and asymptotic properties of solutions*, Ann. Scuola Norm. Sup. Pisa, 10 (1983), pp. 341-355.

[3] D. A. FRANK-KAMENESTSKII, *Diffusion and Heat Transfer in Chemical Kinetics*, Plenum Press, New York, London, 1969.

[4] V. N. IGNAT'EV AND B. G. KUZNETSOV, *A model for the diffusion of a turbulent boundary layer in a polymer*, Čisl. Metody Meh. Splošn. Sredy, 4 (1973), pp. 78-87.

[5] A. V. KAZHIKHOV, *Solvability of the initial and boundary-value problem for the equations of motion of an inhomogeneous viscous incompressible fluid*, Soviet Phys. Dokl., 19 (1974), pp. 331-332.

[6] A. V. KAZHIKHOV AND SH. SMAGULOV, *The correctness of boundary-value problems in a certain diffusion model of an inhomogeneous fluid*, Čisl. Metody Meh. Splošn. Sredy, 7 (1976), pp. 75-92.

[7] ———, *The correctness of boundary-value problems in a diffusion model of an inhomogeneous liquid*, Soviet. Phys. Dokl., 22 (1977), pp. 249-250.

[8] O. A. LADYZHENSKAJA, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.

[9] O. A. LADYZHENSKAJA AND V. A. SOLONNIKOV, *Unique solvability of an initial- and boundary-value problem for viscous incompressible nonhomogeneous fluids*, J. Soviet Math., 9 (1978), pp. 697-749.

[10] J. L. LIONS, *On some problems connected with Navier-Stokes equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978.

[11] ———, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[12] P. SECCHI, *On the initial value problem for the equations of motion of viscous incompressible fluids in the presence of diffusion*, Boll. Un. Mat. Ital. (B), 1 (1982), pp. 1117-1130.

[13] J. SIMON, *Ecoulement d'un fluide non homogène avec une densité initiale s'annulant*, C.R. Acad. Sci. Paris Sér. A-B, 287 (1978), pp. 1009-1012.

[14] SH. SMAGULOV AND K. UTEGENOV, *Asymptotic behavior of the solution of the problem of the flow of an inhomogeneous fluid*, Izv. Akad. Nauk Kazakh. SSR Ser. Fiz.-Mat., 5 (1977), pp. 49-55.

[15] R. TEMAM, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1977.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

003

THE EQUATIONS OF ONE-DIMENSIONAL UNSTEADY FLAME PROPAGATION: EXISTENCE AND UNIQUENESS*

B. LARROUTUROU†

Abstract. This paper deals with the mathematical analysis of a system of partial differential equations describing the time-dependent propagation of a planar flame front within the framework of the well-known isobaric approximation of slow combustion. The problem to be investigated takes the form of a nonlinear mixed initial-boundary value problem in an infinite one-dimensional domain. We show the existence and uniqueness of weak and classical solutions of this problem, depending on the assumptions on the initial data and on the nonlinear temperature dependence of the chemical reaction rates. The crucial point lies in the introduction of a Lagrangian space coordinate, which uncouples the reaction-diffusion equations for the combustion variables from the remaining hydrodynamical subsystem. The analysis then uses some classical arguments of functional analysis, such as the application of the theory of linear semigroups to nonlinear partial differential equations.

Key words. partial differential equations, combustion

AMS(MOS) subject classifications. 35Q20, 76N10, 80A25

1. Introduction. The mathematical analysis of systems of ordinary or partial differential equations arising from the theory of gaseous combustion has received increasing attention in recent years: one can mention for instance several studies of the equations of the stationary planar flame (see [2], [8]) or of the two-dimensional zero Mach number model (see [6]), and in a different domain some mathematical works dealing with the existence and the asymptotic behaviour of the solutions of the Kuramoto–Sivashinsky equation for the flame front instabilities (see [1], [9]).

We present in this paper a new rigorous mathematical result which concerns the time-dependent one-dimensional flame propagation. More precisely, we consider the governing equations of an unsteady planar flame propagating in an infinite channel. These equations, which we recall in § 2, are written using the classical isobaric approximation for reacting flows in open domains (we first consider a simplified one-step chemical mechanism; the extension to chemically complex flames or to nonadiabatic flames is given at the end of the paper). With appropriate hypotheses on the initial data and on the temperature dependence of the reaction rate, we show the global existence and the uniqueness of both weak and classical solutions of the resulting initial-boundary value problem.

The crucial point in our analysis (and in fact the point which restricts our work to the one-dimensional case) lies in the introduction of a Lagrangian space coordinate. This change of coordinates has the effect of decoupling the reaction-diffusion equations for the combustion variables (temperature and mass fraction of the reactant) from the remaining equations for the hydrodynamical variables (density, velocity and pressure). The reactive diffusive system involving the temperature and the mass fraction takes the form of two coupled nonlinear heat equations and is known as the thermodiffusive model for the flame propagation. This parabolic system of partial differential equations is solved in a first step, using classical tools of nonlinear functional analysis such as semigroups generated by linear operators in functional spaces. The remaining subsystem for the hydrodynamical unknowns is then solved in a second step, the temperature

* Received by the editors August 15, 1986; accepted for publication November 3, 1986.
† Institut de Recherche d'Informatique et d'Automatique, Sophia-Antipolis, 06560 Valbonne, France.

being considered given. This provides the existence and uniqueness of solutions of the Lagrangian system. In particular, the analysis shows that no initial data for the hydrodynamical variables need be given for the initial-boundary value problem to be well posed.

Owing to the strong regularizing effect of the heat equation, even a weak solution of the Lagrangian system is continuous. It is then straightforward to come back to the usual Eulerian coordinates, and to prove that similar existence and uniqueness results hold for the original Eulerian system of governing equations.

The paper is organized as follows: (1) Introduction; (2) Governing equations of the flame propagation; (3) Assumptions and main results; (4) Recalling some basic results from semigroup theory; (5) Existence and uniqueness for the combustion variables; (6) Existence and uniqueness for the hydrodynamical variables; (7) Back transformation to the Eulerian variables; (8) Extension to chemically complex flames.

2. Governing equations of the flame propagation.

2.1. Reactive flow equations in one dimension. We are interested in the description of a compressible heat-conducting chemically reacting gaseous mixture with the assumption of a one-dimensional geometry. For the sake of simplicity, we first assume a one-step chemical mechanism $nA \to nB$: the mixture is considered to be made of only two species, the reactant A and the product B. The extension to the case of a chemically complex flame will be investigated in § 8 below.

The reactive gas flow is then described with the usual variables ρ, u, P, T (denoting respectively the total density, velocity, pressure and temperature of the mixture) and an additional variable for the mixture composition, the mass fraction Y of the reactant A (ρY is the separate density of the reactant and $\rho(1-Y)$ is the density of the product). The time-dependent flow of this reactive mixture is then described by the following set of equations (see [5], [7], [14]):

$$
\begin{aligned}
&\rho_\tau + (\rho u)_\xi = 0,\\
&\rho u_\tau + \rho u u_\xi = -P_\xi,\\
&\rho c_p T_\tau + \rho u c_p T_\xi - (\lambda T_\xi)_\xi = Q\omega(Y, T) + P_\tau + uP_\xi,\\
&\rho Y_\tau + \rho u Y_\xi - (\rho D Y_\xi)_\xi = -m\omega(Y, T),\\
&\rho T = \frac{mP}{R},
\end{aligned}
\tag{2.1}
$$

where ξ and τ denote respectively the space and time variables; c_p is the specific heat at constant pressure of the mixture, λ the heat conductivity, D the diffusion coefficient of the reactant A, m its molecular mass and R is the universal gas constant. The effects of viscosity and gravity are neglected. Lastly, $Q(>0)$ is the amount of energy released by the exothermic chemical reaction per unit mass of the reactant, and $\omega(Y, T)$ is the rate at which this reaction proceeds. From the Arrhenius law and the law of mass action, this reaction rate is given by:

$$
\omega(Y, T) = B(T)\left(\frac{\rho Y}{m}\right)^n e^{-E/RT}, \tag{2.2}
$$

where E is the activation energy of the reaction (a constant), and $B(T)$ is some given function of T (which usually has a polynomial-type dependence on T).

2.2. Eulerian form of the flame propagation equations. For writing down the governing equations of the unsteady flame propagation, we will use the so-called

"classical approximation of combustion": the flame propagation is essentially a very subsonic, almost isobaric phenomenon. In other words, the Mach number M of the flow is very small and consequently the pressure variations are also small: $P(\xi, \tau) = P_0 + p(\xi, \tau)$, with $p/P_0 = O(M^2) \ll 1$. For this reason, we may set $P = P_0 = \text{Constant}$ everywhere except in the momentum equation (2.1b) (see [5], [7] for a more detailed discussion of this approximation). The system (2.1) then reduces to

$$\begin{aligned} &\rho_\tau + (\rho u)_\xi = 0, \\ &\rho c_p T_\tau + \rho u c_p T_\xi - (\lambda T_\xi)_\xi = Q\omega(Y, T), \\ &\rho Y_\tau + \rho u Y_\xi - (\rho D Y_\xi)_\xi = -m\omega(Y, T), \\ &\rho T = \frac{mP_0}{R}, \end{aligned} \tag{2.3}$$

$$\rho u_\tau + \rho u u_\xi = -p_\xi. \tag{2.4}$$

Some authors use the system (2.3) alone, replacing the momentum equation (2.4) by $P = P_0$ (see [14]). This is legitimate in one spatial dimension since the only role of the relation (2.4) is the calculation of the small pressure variation p. But this simplification is no more valid when the space dimension N is higher than one, since it eliminates N scalar momentum equations and only one variable p. For this reason we will mainly consider the full system (2.3), (2.4).

The flame propagation equations (2.3), (2.4) will be investigated with the following upstream and downstream boundary conditions:

$$Y(-\infty, t) = Y_u, \quad T(-\infty, t) = T_u, \quad u(-\infty, t) = u^0, \quad p(-\infty, t) = 0 \tag{2.5}$$

(where $Y_u > 0$, $T_u > 0$, $u^0 \in \mathbb{R}$ are given constants) in the fresh mixture, and:

$$Y(+\infty, t) = 0, \qquad T(+\infty, t) = T_b = T_u + \frac{Q}{c_p}\frac{Y_u}{m}, \tag{2.6}$$

in the burnt gases.

2.3. Lagrangian form of the governing equations. From now on we will assume that the Lewis number $\text{Le} = \lambda/\rho c_p D$ and the specific heat c_p are constant. We will also assume that the thermal conductivity of the mixture λ is proportional to the temperature T; this additional assumption will be discussed below, after the derivation of the Lagrangian equations.

We now derive an alternate formulation of the governing equations (2.3), (2.4) using the usual mass-weighted Lagrangian coordinate:

$$x = \int_{\xi(0,t)}^{\xi(x,t)} \rho(\xi', t)\, d\xi'. \tag{2.7}$$

Although the use of this transformation is classical, we detail the calculation for sake of completeness. Let us define a Lagrangian coordinate (i.e. a variable whose value, defined at time $\tau = 0$, remains constant during the flow for each fluid particle) by setting:

$$x = \int_0^{\xi} \rho(\xi', 0)\, d\xi'.$$

We also set $t = \tau$. Then $x(\xi, \tau)$ represents the Lagrangian coordinate of the particle which is located at the abscissa ξ at time τ and the last relation is to be read as

$x(\xi, 0) = \int_0^{\xi} \rho(\xi', 0)\, d\xi'$. Inversely, $\xi(x, t)$ is the position at time t of the fluid particle whose Lagrangian coordinate is x. Therefore we have, by definition:

$$\xi_t = u \quad \text{or} \quad \frac{\partial}{\partial t}\xi(x, t) = u[\xi(x, t), t].$$

We can then write:

$$\begin{aligned}\frac{d}{dt}\left[\int_{\xi(0,t)}^{\xi(x,t)} \rho(\xi', t)\, d\xi'\right] &= \frac{\partial \xi}{\partial t}(x, t)\rho[\xi(x, t), t] - \frac{\partial \xi}{\partial t}(0, t)\rho[\xi(0, t), t] \\ &\quad + \int_{\xi(0,t)}^{\xi(x,t)} \frac{\partial \rho}{\partial t}(\xi', t)\, d\xi' \\ &= (\rho u)[\xi(x, t), t] - (\rho u)[\xi(0, t), t] \\ &\quad + \int_{\xi(0,t)}^{\xi(x,t)} \frac{\partial \rho}{\partial t}(\xi', t)\, d\xi' \\ &= \int_{\xi(0,t)}^{\xi(x,t)} [\rho_t + (\rho u)_\xi](\xi', t)\, d\xi' = 0,\end{aligned}$$

whence:

$$\int_{\xi(0,t)}^{\xi(x,t)} \rho(\xi', t)\, d\xi' = \int_{\xi(0,0)}^{\xi(x,0)} \rho(\xi', 0)\, d\xi' = x,$$

which is exactly (2.7).

Differentiating (2.7) with respect to x gives:

$$1 = \rho \xi_x \quad \text{or} \quad \frac{\partial}{\partial x}\xi(x, t) = \frac{1}{\rho[\xi(x, t), t]}.$$

We then have in matrix form (writing simply $u(x, t)$ for $u[\xi(x, t), t]$):

$$\begin{pmatrix} \xi_x & \xi_t \\ \tau_x & \tau_t \end{pmatrix} = \begin{pmatrix} \rho^{-1} & u \\ 0 & 1 \end{pmatrix},$$

which implies:

$$\begin{pmatrix} x_\xi & x_\tau \\ t_\xi & t_\tau \end{pmatrix} = \begin{pmatrix} \rho & -\rho u \\ 0 & 1 \end{pmatrix}. \tag{2.8}$$

Remark 2.1. The mass balance equation (2.3a) has been crucial for introducing the new variable x. This amounts to noticing that a variable X satisfying $X_\xi = \rho$, $X_\tau = -\rho u$ (i.e. (2.8)) could have been introduced directly, since (2.3a) insures that $(\partial/\partial\tau)(X_\xi) = (\partial/\partial\xi)(X_\tau)$.

We can now derive the Lagrangian form of the flame propagation equations. For any quantity F we have $F_\tau = F_t - \rho u F_x$, $F_\xi = \rho F_x$, and the system (2.3), (2.4) becomes

$$\begin{aligned} &\rho_t + \rho^2 u_x = 0, \\ &u_t + p_x = 0, \\ &T_t = \frac{Q}{c_p}\frac{\omega}{\rho} + \frac{1}{c_p}(\lambda \rho T_x)_x, \\ &Y_t = -m\frac{\omega}{\rho} + (\rho^2 D Y_x)_x, \\ &\rho T = \frac{m P_0}{R}. \end{aligned} \tag{2.9}$$

To nondimensionalize these equations, we refer the mass fraction to $Y_0 = Y_u$, the temperature to $T_0 = T_b - T_u = (Q/c_p)(Y_u/m)$, the density to $\rho_0 = mP_0/RT_0$. Denoting $(\lambda\rho)_0$ the constant value of $\lambda\rho = (mP_0/R)(\lambda/T)$, we relate the time unit t_0 and the "Lagrangian unit" x_0 by: $x_0^2 = t_0(\lambda\rho)_0/c_p$. The velocity is then referred to $u_0 = x_0/\rho_0 t_0$ and the pressure variation to $p_0 = \rho_0 u_0^2$.

Setting $\Theta = T - T_u$ and denoting by $\hat{\Theta}$, $\hat{Y}$, $\hat{\rho}$, $\hat{u}$, $\hat{p}$ the nondimensionalized variables, we obtain the following expressions for the Lagrangian equations (2.9) and boundary conditions (2.5), (2.6):

$$\hat{\Theta}_t = \hat{\Theta}_{xx} + \Omega(\hat{Y}, \hat{\Theta}), \qquad \hat{Y}_t = \frac{1}{\mathrm{Le}}\hat{Y}_{xx} - \Omega(\hat{Y}, \hat{\Theta}), \tag{2.10}$$

$$(\hat{\Theta} + \alpha)\hat{\rho} = 1, \quad \hat{u}_x = \hat{\Theta}_t, \quad \hat{u}_t + \hat{p}_x = 0, \tag{2.11}$$

$$\hat{\Theta}(-\infty, t) = 0, \qquad \hat{\Theta}(+\infty, t) = 1,$$

$$\hat{Y}(-\infty, t) = 1, \qquad \hat{Y}(+\infty, t) = 0, \tag{2.12}$$

$$\hat{u}(-\infty, t) = \hat{u}^0, \qquad \hat{p}(-\infty, t) = 0,$$

where x and t now represent the nondimensionalized Lagrangian coordinates, $\alpha = T_u/(T_b - T_u)$ is a nondimensional heat release parameter, and $\Omega(\hat{Y}, \hat{\Theta}) = (Q/c_p)(R/mP_0)t_0(\omega/\hat{\rho})$ is the normalized reaction rate. In the sequel, we will assume using (2.2) that Ω is given by:

$$\Omega(\hat{Y}, \hat{\Theta}) = \hat{Y}^n f(\hat{\Theta}),$$

where f is a positive continuous function satisfying $f(0) = 0$.

Remark 2.2. The assumption $f(0) = 0$ is not fulfilled in view of the expression (2.2) of the reaction rate ω since $e^{-E/RT_u} \neq 0$. This is the well-known "cold boundary difficulty," on which a lot has already been said (see [5]). Let us just point out that this hypothesis is necessary for the mathematical problem (2.10), (2.12) to be well posed.

It should be emphasized here that the use of the Lagrangian coordinate (2.7) uncouples the equations (2.10) for the combustion field $(\hat{\Theta}, \hat{Y})$ (which take the form of a purely diffusive reaction system) from the equations (2.11) for the hydrodynamical variables $(\hat{\rho}, \hat{u}, \hat{p})$. Moreover, the form of these hydrodynamical equations leads one to think that no initial data for the density, velocity or pressure is needed to determine the profiles of these variables at positive time values: these hydrodynamical profiles $\hat{\rho}(\cdot, t)$, $\hat{u}(\cdot, t)$, $\hat{p}(\cdot, t)$ for $t > 0$ only depend on the temperature profiles $\hat{\Theta}(\cdot, t')$ for $t' \geqq 0$; we first have to study the nonlinear parabolic system (2.10), and (2.11) will be investigated in a second step.

Remark 2.3. The assumption $\lambda\rho = \text{Constant}$, or $\lambda/T = \text{Constant}$ only affects the expression of the diffusive terms in the temperature and mass fraction equations: these terms take the form $\hat{\Theta}_{xx}$ and $(1/\mathrm{Le})\hat{Y}_{xx}$ instead of $[(\lambda\rho)T_x]_x$ and $(1/\mathrm{Le})[(\lambda\rho)Y_x]_x$ where, in complete generality, $\lambda\rho$ is a function of T and Y. Nevertheless, it can be noticed that this hypothesis (which is rather classical in combustion theory, see [11]) does not change the preceding remarks about the nature of the Lagrangian system (2.10), (2.11). We hope to extend our mathematical analysis to the case of a nonconstant λ/T ratio in a forthcoming paper.

Remark 2.4. In the classical nondimensionalization of the Eulerian equations (2.3), (2.4) (see [5], [7]), the length and time scales ξ_0 and $\tau_0 = t_0$ are related to the thermal diffusion coefficient $(\lambda/\rho c_p)_0 = (\lambda\rho)_0/\rho_0^2 c_p$ and to the velocity unit u_0 by the

identities:

$$\frac{\xi_0^2}{\tau_0}=\frac{(\lambda\rho)_0}{\rho_0^2 c_p} \quad \text{and} \quad u_0=\frac{\xi_0}{\tau_0}.$$

In our case, the units used above to nondimensionalize the equations (2.9) have essentially been chosen in order to simplify the Lagrangian system (2.10), (2.11), which will play a crucial role in the sequel. Therefore, these units are not quite usual, and the above relations are replaced by:

$$\frac{\xi_0^2}{\tau_0}\left[\int_0^1 \hat{\rho}(\zeta,0)d\zeta\right]^2=\frac{(\lambda\rho)_0}{\rho_0^2 c_p} \quad \text{and} \quad u_0=\frac{\xi_0}{\tau_0}\int_0^1 \hat{\rho}(\zeta,0)\,d\zeta,$$

since $x_0=\int_0^{\xi_0}\rho(\xi',0)\,d\xi'=\rho_0\xi_0\int_0^1\hat{\rho}(\zeta,0)\,d\zeta$.

3. Assumptions and main results.

3.1. Statement of the problem. The aim of this paper is to investigate the following version of (2.10)-(2.12):

$$\begin{aligned}
&\Theta_t-\Theta_{xx}=\Omega(Y,\Theta)=Y^n f(\Theta),\\
&Y_t-\frac{Y_{xx}}{\mathrm{Le}}=-\Omega(Y,\Theta),\\
&(\Theta+\alpha)\rho=1, \qquad (3.1)\\
&u_x=\Theta_t \quad \text{for } x\in\mathbb{R},\ t\in\mathbb{R}_+,\\
&\Theta(x,0)=\Theta_0(x), \qquad Y(x,0)=Y_0(x),\\
&\Theta(-\infty,t)=0, \qquad \Theta(+\infty,t)=1,\\
&Y(-\infty,t)=1, \qquad Y(+\infty,t)=0, \qquad (3.2)\\
&u(-\infty,t)=u^0,\\
&u_t+p_x=0, \qquad p(-\infty,t)=0. \qquad (3.3)
\end{aligned}$$

We will also study the corresponding normalized Eulerian formulation in conservative form:

$$\begin{aligned}
&\rho_\tau+(\rho u)_\xi=0,\\
&(\rho u)_\tau+(\rho u^2)_\xi+p_\xi=0,\\
&(\rho\Theta)_\tau+(\rho u\Theta)_\xi-\left(\frac{\Theta_\xi}{\rho}\right)_\xi=\rho\Omega(Y,\Theta),\\
&(\rho Y)_\tau+(\rho u Y)_\xi-\frac{1}{\mathrm{Le}}\left(\frac{Y_\xi}{\rho}\right)_\xi=-\rho\Omega(Y,\Theta), \qquad (3.4)\\
&(\Theta+\alpha)\rho=1 \qquad \text{for } x\in\mathbb{R},\ t\in\mathbb{R}_+,\\
&\Theta(x,0)=\Theta_0(x), \qquad Y(x,0)=Y_0(x),\\
&\Theta(-\infty,t)=0, \qquad \Theta(+\infty,t)=1,\\
&Y(-\infty,t)=1, \qquad Y(+\infty,t)=0, \qquad (3.5)\\
&u(-\infty,t)=u^0, \qquad p(-\infty,t)=0.
\end{aligned}$$

It can be noticed here that initial data are prescribed only for the temperature and mass fraction (Θ, Y) and not for the hydrodynamical unknowns (ρ, u, p).

For the investigation of these two problems, we will mainly focus on two types of solutions, which we define precisely below.

DEFINITION 3.1. (Θ, Y, ρ, u, p) is a *weak solution* of problem (3.1)-(3.3) if the three following properties hold:

(1) $(\Theta, Y, \rho, u, p) \in [L^{\infty}_{\text{loc}}(\mathbb{R}\times\mathbb{R}_+)]^5$ and (Θ, Y, ρ, u, p) is a solution of (3.1)-(3.3a) in the sense of the distributions:

$$\int_{\mathbb{R}\times\mathbb{R}_+} [-\Theta\eta_t + \Theta\eta_{xx} - \Omega\eta] = \int_{\mathbb{R}} \Theta_0\eta(\cdot, 0),$$

$$\int_{\mathbb{R}\times\mathbb{R}_+} [-Y\eta_t + \frac{1}{\text{Le}} Y\eta_{xx} + \Omega\eta] = \int_{\mathbb{R}} Y_0\eta(\cdot, 0),$$

$$\int_{\mathbb{R}\times\mathbb{R}_+} [(\Theta+\alpha)\rho\eta - \eta] = 0,$$

$$\int_{\mathbb{R}\times\mathbb{R}_+} [u\eta_x - \Theta\eta_t] = \int_{\mathbb{R}} \Theta_0\eta(\cdot, 0),$$

$$\int_{\mathbb{R}\times\mathbb{R}_+} -[u\eta_t + p\eta_x] = \int_{\mathbb{R}} u(\cdot, 0)\eta(\cdot, 0) \quad \text{for any } \eta \in D(\mathbb{R}\times\mathbb{R}_+).$$

(2) The boundary conditions (3.2) hold in the classical sense for $t > 0$ and (3.3b) holds in the following weak sense:

$$\forall t > 0, \quad \exists p_1 \in L^2(\mathbb{R}), \quad \lim_{x\to-\infty} [p(x, t) - p_1(x)] = 0. \tag{3.6}$$

(3) The following inequalities (which are necessary from a physical standpoint) hold:

$$\Theta(x, t) \geqq 0, \quad 0 \leqq Y(x, t) \leqq 1 \quad \text{a.e. on } \mathbb{R}\times\mathbb{R}_+.$$

Moreover, $\Theta \in L^{\infty}_{\text{loc}}(\mathbb{R}_+, L^{\infty}(\mathbb{R}))$.

DEFINITION 3.2. A *weak solution* (Θ, Y, ρ, u, p) of problem (3.1)-(3.3) is a *smooth solution* if and only if:

(1) All the functions and all the partial derivatives appearing in the equations (3.1) and (3.3a) are continuous with respect to both variables x and t on $\mathbb{R}\times\mathbb{R}_+$;

(2) The boundary conditions (3.2) and (3.3b) are fulfilled in the classical sense for $t \geqq 0$.

Similar definitions hold for the solutions of (3.4), (3.5).

3.2. Assumptions and notation. Before stating the main hypotheses which will be used for investigating the two above problems, we need to introduce two functions γ and γ_1 of $C^{\infty}(\mathbb{R})$ satisfying

$$\begin{aligned} &\gamma = 0 \quad \text{on } (-\infty, -1], \quad 0 \leqq \gamma \leqq 1 \quad \text{on } [-1, 1], \quad \gamma = 1 \quad \text{on } [1, +\infty), \\ &\gamma_1 = 1 - \gamma. \end{aligned} \tag{3.7}$$

We will set:

$$\varphi_0(x) = \Theta_0(x) - \gamma(x), \qquad \psi_0(x) = Y_0(x) - \gamma_1(x). \tag{3.8}$$

The following assumptions will be used in the theorems stated below:

$$\begin{cases} \varphi_0 \in L^2(\mathbb{R}) \cap L^1(\mathbb{R}), \qquad \psi_0 \in L^2(\mathbb{R}) \cap L^1(\mathbb{R}); \\ \Theta_0 \in L^\infty(\mathbb{R}), \quad \Theta_0(x) \geqq 0 \quad \text{a.e.}, \\ Y_0(x) \in [0,1] \quad \text{a.e.}, \\ \mathrm{Le} > 0 \quad \text{and} \quad n \in \mathbb{N}^* \text{ are given}, \\ f \in C(\mathbb{R}_+, \mathbb{R}_+), \qquad f(0) = 0, \\ \forall \vartheta > 0, \quad f \text{ is Lipschitz-continuous on } [0, \vartheta]. \end{cases} \tag{3.9}$$

Moreover, we will sometimes need some of the following more technical hypotheses:

$$\exists C_f > 0, \quad \forall \vartheta \in \mathbb{R}_+, \qquad |f(\vartheta)| \leqq C_f |\vartheta|, \tag{3.10}$$

$$\varphi_0 \in H^2(\mathbb{R}), \qquad \psi_0 \in H^2(\mathbb{R}), \tag{3.11}$$

$$f \in C^1(\mathbb{R}_+, \mathbb{R}_+),$$

$$\exists \beta > \frac{1}{2}, \quad \overline{\lim_{\vartheta \to 0}} \frac{|f'(\vartheta)|}{\vartheta^\beta} < +\infty, \tag{3.12}$$

$$\exists \mu > \frac{3}{2\beta}, \quad \sup_{x \in \mathbb{R}_-} \Theta_0(x) |x|^\mu < +\infty, \tag{3.13}$$

$$\varphi_0 \in H^4(\mathbb{R}), \qquad \psi_0 \in H^4(\mathbb{R}), \tag{3.14}$$

$$\begin{aligned} &f \in C^2(\mathbb{R}_+, \mathbb{R}_+), \\ &\forall \vartheta > 0, \quad f_{xx} \text{ is Lipschitz-continuous on } [0, \vartheta]. \end{aligned} \tag{3.15}$$

From now on, we will denote $L^p = L^p(\mathbb{R})$, for $p \in [1, +\infty)$, and $\|\varphi\|_p = \|\varphi\|_{L^p}$ or $\|(\varphi, \psi)\|_p = \max(\|\varphi\|_{L^p}, \|\psi\|_{L^p})$. Furthermore, for $m \in \mathbb{N}^*$, we set $H^m = H^m(\mathbb{R}) = W^{m,2}(\mathbb{R})$.

3.3. Results concerning the Lagrangian formulation. The first of our theorems deals with the problem (3.1), (3.2) without the pressure variable.

THEOREM 3.3. *Assume that the hypotheses* (3.9) *and* (3.10) *hold. Then there exists a unique* weak solution (Θ, Y, ρ, u) *of* (3.1), (3.2) *in* $\mathbb{R} \times \mathbb{R}_+$ *satisfying*:

$$\Theta - \gamma, \; Y - \gamma_1 \in C(\mathbb{R}_+, L^2). \tag{3.16}$$

Furthermore, this solution satisfies:

$$\begin{aligned} &\Theta, Y, \rho \in C(\mathbb{R}_+, L^\infty) \cap C(\mathbb{R}_+^*, C^1(\mathbb{R})), \\ &\Theta - \gamma, \; Y - \gamma_1 \in C^1(\mathbb{R}_+^*, L^2), \\ &u \in C(\mathbb{R}_+^*, C(\mathbb{R}) \cap L^\infty). \end{aligned} \tag{3.17}$$

Concerning the complete system (3.1)-(3.3), we have the following two results.

THEOREM 3.4. *Assume that the hypotheses* (3.9)-(3.13) *hold. Then there exists a unique weak solution* (Θ, Y, ρ, u, p) *of* (3.1)-(3.3) *in* $\mathbb{R}\times\mathbb{R}_+$ *satisfying* (3.16). *Moreover this solution satisfies* (3.17).

THEOREM 3.5. *Assume that all the hypotheses* (3.9) *to* (3.15) *hold. Then there exists a unique smooth solution of* (3.1)-(3.3) *in* $\mathbb{R}\times\mathbb{R}_+$. *This solution satisfies*:

$$\begin{aligned} &\Theta, Y, \rho \in C(\mathbb{R}_+, C^3(\mathbb{R})) \cap C^1(\mathbb{R}_+, C^1(\mathbb{R})), \\ &u \in C(\mathbb{R}_+, C^2(\mathbb{R})) \cap C^1(\mathbb{R}_+, C(\mathbb{R})). \end{aligned} \tag{3.18}$$

These three theorems will be proved in §§ 5 and 6 below.

3.4. Results concerning the Eulerian formulation. Analogous results hold for the Eulerian problem (3.4), (3.5).

THEOREM 3.6. *Assume that the hypotheses* (3.9)-(3.13) *hold. Then there exists a unique weak solution* (Θ, Y, ρ, u, p) *of* (3.4), (3.5) *in* $\mathbb{R}\times\mathbb{R}_+$ *satisfying* (3.16) *and*:

$$u, \rho \in C(\mathbb{R}\times\mathbb{R}_+, \mathbb{R}).$$

Moreover, this solution satisfies (3.17).

THEOREM 3.7. *Assume that all the hypotheses* (3.9) *to* (3.15) *hold. Then there exists a unique smooth solution of* (3.4), (3.5) *in* $\mathbb{R}\times\mathbb{R}_+$. *This solution satisfies* (3.18).

The proof of these two last results is detailed in § 7.

4. Recalling some basic results from semigroup theory. In this section, we briefly recall some classical results from functional analysis which will be needed in the following sections. We refer the reader to [3], [4], [10], [15] for more details and for the proofs of these results.

4.1. Semigroups of linear operators. Let us first recall some basic definitions and results about maximal monotone linear operators.

Let H be a real Hilbert space and A be an unbounded linear operator defined on the subspace $D(A)\subset H$. The operator A is said to be maximal monotone if and only if:

$$\begin{aligned} &\forall u \in D(A), \quad (Au, u) \geqq 0, \\ &\forall v \in H, \quad \exists u \in D(A), \quad v = u + Au. \end{aligned} \tag{4.1}$$

The basic property is the theorem of Hille and Yosida.

THEOREM 4.1 (Hille and Yosida). *Let H be a real Hilbert space and A be a maximal monotone linear operator defined on the subspace $D(A)\subset H$. For $u_0\in D(A)$, the problem*:

$$\begin{aligned} &\frac{du}{dt} + Au = 0 \quad \text{for } t \geqq 0, \\ &u(0) = u_0 \end{aligned} \tag{4.2}$$

has a unique solution in $C(\mathbb{R}_+, D(A)) \cap C^1(\mathbb{R}_+, H)$.

Let $u(t)$ be the solution of (4.2) for $t\geqq 0$; we set $u(t) = R(t)u_0$, where $R(t)$ is a linear operator from $D(A)$ into H. Since it follows from (4.1) that $D(A)$ is a dense subspace of H, we can extend $R(t)$ to the whole space H; the resulting operator, which we still denote by $R(t)$, is (by definition) the linear semigroup generated by $-A$.

Let us finally recall that a maximal monotone operator A is self-adjoint if and only if, for all $(u, v)\in D(A)^2$, $(Au, v) = (u, Av)$.

4.2. Nonlinear equations. We are going to consider some problems of the form:

$$\begin{aligned} &\frac{du}{dt} + Au = F(u) \quad \text{for } t \geqq 0, \\ &u(0) = u_0, \end{aligned} \tag{4.3}$$

where A is a linear self-adjoint maximal monotone operator, $u_0 \in H$ and $F \in C(H, H)$. Before stating results about the existence of a solution of this problem, we specify which type of solution will be considered.

DEFINITION 4.2. u is a *classical solution* of (4.3) on an interval $[0, T)$ if and only if u satisfies (4.3) in the classical sense, i.e., with:

$$u \in C^1([0, T), H) \cap C([0, T), D(A)).$$

u is a *weak solution* of (4.3) on $[0, T)$ if and only if $u \in C([0, T), H)$ and:

$$\forall t \in [0, T), \quad u(t) = R(t)u_0 + \int_0^t R(t-s)F[u(s)]\, ds. \tag{4.4}$$

We can then state the following two theorems.

THEOREM 4.3. *Let H be a real Hilbert space and A be a linear self-adjoint maximal monotone operator defined on the subspace $D(A) \subset H$. Assume that F is a Lipschitz-continuous mapping from H into itself. Then for any $u_0 \in H$, there exists a unique weak solution of* (4.3) *in $\mathbb{R}_+$ and this solution u is classical on $\mathbb{R}_+^*$.*

Moreover, if $u_0 \in D(A)$, then u is a classical solution on $\mathbb{R}_+$.

THEOREM 4.4. *Let H be a real Hilbert space and A be a linear self-adjoint maximal monotone operator defined on the subspace $D(A) \subset H$. Assume that F is a Lipschitz-continuous mapping from any bounded subset of H into H. Then for any $u_0 \in H$, there exists $T_{\max} > 0$ such that a unique weak solution of* (4.3) *exists on $[0, T_{\max})$; this solution u is classical on $(0, T_{\max})$ and the following alternative holds:*

$$\begin{aligned} &\textit{Either:} \quad T_{\max} = +\infty, \\ &\textit{Or:} \quad \lim_{t \to T_{\max}} \|u(t)\|_H = +\infty. \end{aligned}$$

Moreover, if $u_0 \in D(A)$, then u is a classical solution on $[0, T_{\max})$.

4.3. Application to the heat equation. We now consider the case $H = L^2$, and the operator:

$$A: \begin{vmatrix} D(A) = H^2 \to L^2, \\ \varphi \to -\varphi_{xx}. \end{vmatrix}$$

Problem (4.3) then becomes a nonlinear heat equation; Theorems 4.3 and 4.4 apply to this case because of the following lemma.

LEMMA 4.5. *A is a self-adjoint maximal monotone operator.*

Let $S(t)$ be the semigroup generated by $-A$; the following properties of this semigroup will be useful in the sequel.

LEMMA 4.6. *The following properties hold for the semigroup $S(t)$:*

$$\forall p \in [1, \infty), \quad \forall \varphi \in L^2 \cap L^p, \quad \forall t \in \mathbb{R}_+, \quad \|S(t)\varphi\|_{L^p} \leqq \|\varphi\|_{L^p},$$
$$\forall \varphi \in L^2 \cap L^\infty, \quad S(\cdot)\varphi \in C[\mathbb{R}_+, L^\infty(\mathbb{R})].$$

LEMMA 4.7. *Let $u_0 \in L^2$. The following explicit expression holds for $S(t)u_0$:*

$$[S(t)u_0](x) = \frac{1}{\sqrt{4\pi t}} \int_{\mathbb{R}} u_0(y)\, e^{|x-y|^2/4t}\, dy. \tag{4.5}$$

5. Existence and uniqueness for the combustion variables.

5.1. Statement of the problem and main results. The aim of this section is to study the subsystem of the reaction-diffusion equations for the temperature and mass fraction:

$$\Theta_t - \Theta_{xx} = \Omega(Y, \Theta) = Y^n f(\Theta), \qquad Y_t - \frac{Y_{xx}}{\mathrm{Le}} = -\Omega(Y, \Theta) \quad \text{for } x \in \mathbb{R},\ t \in \mathbb{R}_+,$$

(5.1)

$$\Theta(x, 0) = \Theta_0(x), \qquad Y(x, 0) = Y_0(x),$$

$$\Theta(-\infty, t) = 0, \qquad \Theta(+\infty, t) = 1,$$

(5.2)

$$Y(-\infty, t) = 1, \qquad Y(+\infty, t) = 0.$$

Before stating the results concerning the existence and uniqueness of a solution of problem (5.1), (5.2) we introduce a new formulation of this problem. In order to apply some of the results recalled in the preceding section, we define new unknowns (φ, ψ) satisfying zero boundary condition; we therefore use the functions γ and γ_1 introduced in (3.7), define (φ_0, ψ_0) as in (3.8) and set:

$$\varphi(x, t) = \Theta(x, t) - \gamma(x), \qquad \psi(x, t) = Y(x, t) - \gamma_1(x). \tag{5.3}$$

Finally we extend the domain of definition of f by setting: $f \equiv 0$ on $\mathbb{R}_-$, and we define g by:

$$g(\xi) = \begin{cases} 0 & \text{if } \xi \leqq 0, \\ \xi^n & \text{if } \xi \geqq 0. \end{cases} \tag{5.4}$$

The system (5.1), (5.2) can now be rewritten as:

$$\varphi_t - \varphi_{xx} = f(\varphi + \gamma) g(\psi + \gamma_1) + \gamma_{xx},$$

$$\psi_t - \frac{\psi_{xx}}{\mathrm{Le}} = -f(\varphi + \gamma) g(\psi + \gamma_1) - \frac{\gamma_{xx}}{\mathrm{Le}}, \tag{5.5}$$

$$\varphi(x, 0) = \varphi_0(x), \qquad \psi(x, 0) = \psi_0(x),$$

$$\varphi(-\infty, t) = \varphi(+\infty, t) = \psi(-\infty, t) = \psi(+\infty, t) = 0. \tag{5.6}$$

The next lemma shows that problem (5.5) does belong to the general framework of the preceding section. Consider the linear operator:

$$A: \begin{vmatrix} D(A) = H^2 \times H^2 \to L^2 \times L^2, \\ (\varphi, \psi) \to \left(-\varphi_{xx}, -\frac{\psi_{xx}}{\mathrm{Le}} \right). \end{vmatrix}$$

We then have the following lemma.

LEMMA 5.1. *A is a maximal monotone self-adjoint operator.*

Proof. The proof is obvious from Lemma 4.5. □

Remark 5.2. Let S^2 be the continuous linear semigroup generated by $-A$. The two following properties follow easily from Lemma 4.6:

$$\forall p \in [1, \infty), \quad \forall (\varphi, \psi) \in L^2 \times L^2 \cap L^p \times L^p, \quad \forall t \in \mathbb{R}_+, \quad \| S^2(t)(\varphi, \psi) \|_p \leqq \| (\varphi, \psi) \|_p,$$

$$\forall (\varphi, \psi) \in L^2 \times L^2 \cap L^\infty \times L^\infty, \quad S^2(\cdot)(\varphi, \psi) \in C(\mathbb{R}_+, L^\infty \times L^\infty).$$

We can now make precise what the solutions of (5.5) may be, in view of Definitions 3.2 and 4.2. For the more general problem:

$$\begin{aligned} &\varphi_t - \varphi_{xx} = h_1(\varphi, \psi, x), \\ &\psi_t - \frac{\psi_{xx}}{\mathrm{Le}} = h_2(\varphi, \psi, x), \\ &\varphi(x, 0) = \varphi_0(x), \qquad \psi(x, 0) = \psi_0(x), \end{aligned} \tag{5.7}$$

we state the following.

DEFINITION 5.3. (φ, ψ) is a *weak solution* of (5.7) on $\mathbb{R}\times[0, T)$ if and only if:

$$\varphi, \psi \in C([0, T), L^2), \qquad H(\varphi, \psi) \in C([0, T), L^2 \times L^2),$$

$$\forall t \in [0, T), \quad (\varphi, \psi)(t) = S^2(t)(\varphi_0, \psi_0) + \int_0^t S^2(t-s) H[(\varphi, \psi)(s)]\, ds,$$

where $H(\varphi, \psi) = [h_1(\varphi, \psi, x), h_2(\varphi, \psi, x)]$.

A *weak solution* (φ, ψ) of (5.7) is a *classical solution* on the interval K of $\mathbb{R}_+$ if and only if:

$$\varphi, \psi \in C^1(K, L^2) \cap C(K, H^2).$$

A classical solution (φ, ψ) of (5.7) is a *smooth solution* if and only if (φ, ψ) and all the partial derivatives appearing in (5.7) are continuous with respect to both variables x and t.

DEFINITION 5.4. (Θ, Y) is a *weak* (resp., *classical, smooth*) solution of (5.1) on $\mathbb{R}\times[0, T)$ if and only if (Θ, Y) is related to a *weak* (resp., *classical, smooth*) solution (φ, ψ) of (5.5) on $\mathbb{R}\times[0, T)$ by (5.3), and satisfies:

$$\Theta \in L^\infty_{\mathrm{loc}}(\mathbb{R}_+, L^\infty),$$

$$\Theta(x, t) \geqq 0, \qquad 0 \leqq Y(x, t) \leqq 1 \quad \text{a.e. on } \mathbb{R}\times[0, T).$$

Remark 5.5. It is easily checked that a *weak solution* (φ, ψ) of (5.5) on $\mathbb{R}\times\mathbb{R}_+$ which is also a *classical solution* on $\mathbb{R}_+^*$ is a solution in the sense of the distributions. Let indeed $\eta \in D(\mathbb{R}\times\mathbb{R}_+)$. Assuming that $\mathrm{Supp}\,(\eta) \subset (-M, M)\times[0, T)$, we set: $K = (-M, M)\times[0, T)$ and $K_\varepsilon = (-M, M)\times(\varepsilon, T)$. Since (φ, ψ) is a classical solution on $\mathbb{R}\times K_\varepsilon$, φ, ψ and η are in $H^1(K_\varepsilon)$. We can then apply Green's formula to get:

$$\int_{K_\varepsilon} [-\varphi\eta_t - (\varphi+\gamma)\eta_{xx} - f(\varphi+\gamma)g(\psi+\gamma_1)\eta] = \int_{-M}^{M} \varphi(x, \varepsilon)\eta(x, \varepsilon)\, dx.$$

As $\varphi \in C([0, T), L^2)$, we can take the limit $\varepsilon \to 0$ in the last relation to get:

$$\int_K [-\varphi\eta_t - (\varphi+\gamma)\eta_{xx} - f(\varphi+\gamma)g(\psi+\gamma_1)\eta] = \int_{-M}^{M} \varphi_0(x)\eta(x, 0)\, dx,$$

which (together with the analogous relation for ψ) shows that (φ, ψ) is a solution of (5.5) in the sense of $D'(\mathbb{R}\times\mathbb{R}_+)$.

In the same way, a *weak solution* (Θ, Y) of (5.1) is a solution in the sense of distributions.

We are now ready to state the main results about problems (5.1), (5.2) and (5.5), (5.6). For the sake of simplicity, we are using both the new unknowns (φ, ψ) and the old ones (Θ, Y).

THEOREM 5.6. *Under the hypotheses* (3.9) *and* (3.10), *there exists a unique solution* (Θ, Y) *of problem* (5.1), (5.2). *The corresponding solution* (φ, ψ) *of* (5.5), (5.6) *satisfies*:

$$\varphi, \psi \in C(\mathbb{R}_+, L^2 \cap L^\infty) \cap C^1(\mathbb{R}_+^*, L^2) \cap C(\mathbb{R}_+^*, H^2).$$

COROLLARY 5.7. *Under the hypotheses* (3.9), (3.10) *and* (3.15), *the solution* (Θ, Y) *of* (5.1), (5.2) *is a smooth solution on* $\mathbb{R}\times\mathbb{R}_+^*$.

Moreover, if $\varphi_0, \psi_0 \in H^4$, (Θ, Y) *is a smooth solution on* $\mathbb{R}\times\mathbb{R}_+$.

5.2. A lemma for systems of type (5.5). We begin the proof of the above theorems with the next result, which will be used several times in the sequel.

LEMMA 5.8. *Let* f_0 *and* g_0 *be two bounded Lipschitz-continuous functions on* $\mathbb{R}$, *with* $f_0(0)=0$, $g_0(0)=0$; *consider the problem*:

$$\begin{aligned} &\varphi_t-\varphi_{xx}=f_0(\varphi+\gamma)g_0(\psi+\gamma_1)+\gamma_{xx},\\ &\psi_t-\frac{\psi_{xx}}{\mathrm{Le}}=-f_0(\varphi+\gamma)g_0(\psi+\gamma_1)-\frac{\gamma_{xx}}{\mathrm{Le}},\\ &\varphi(x,0)=\varphi_0(x),\qquad \psi(x,0)=\psi_0(x). \end{aligned} \tag{5.8}$$

For any $(\varphi_0, \psi_0)\in L^2\times L^2$, *the problem* (5.8) *has a unique solution* (φ, ψ) *in* $C(\mathbb{R}_+, L^2\times L^2)$; *this solution is a classical solution on* $\mathbb{R}_+^*$:

$$\varphi, \psi \in C(\mathbb{R}_+^*, H^2)\cap C^1(\mathbb{R}_+^*, L^2).$$

Proof. Define the mapping F_0 by:

$$F_0(\varphi,\psi)=\left[f_0(\varphi+\gamma)g_0(\psi+\gamma_1)+\gamma_{xx}, -f_0(\varphi+\gamma)g_0(\psi+\gamma_1)-\frac{\gamma_{xx}}{\mathrm{Le}}\right]$$

for $\varphi, \psi \in L^2$. In view of Theorem 4.3, it suffices to show that F_0 is a Lipschitz-continuous mapping from $L^2\times L^2$ into itself.

Let $h=f_0(\varphi+\gamma)g_0(\psi+\gamma_1)$. It is classical to show that $h\in L^2$ when $\varphi, \psi\in L^2$. Let us simply check that h is Lipschitz-continuous from $L^2\times L^2$ into L^2. Let M_f, M_g, L_f, L_g be real constants such that:

$$\forall \xi\in\mathbb{R},\quad |f_0(\xi)|\leqq M_f,\quad |g_0(\xi)|\leqq M_g,$$

$$\forall(\xi,\eta)\in\mathbb{R}^2,\quad |f_0(\xi)-f_0(\eta)|\leqq L_f|\xi-\eta|,\quad |g_0(\xi)-g_0(\eta)|\leqq L_g|\xi-\eta|.$$

For $\varphi_1, \psi_1\in L^2$, $\varphi_2, \psi_2\in L^2$, we have:

$$\begin{aligned} h_1-h_2&=f_0(\varphi_1+\gamma)g_0(\psi_1+\gamma_1)-f_0(\varphi_2+\gamma)g_0(\psi_2+\gamma_1)\\ &=f_0(\varphi_1+\gamma)[g_0(\psi_1+\gamma_1)-g_0(\psi_2+\gamma_1)]+g_0(\psi_2+\gamma_1)[f_0(\varphi_1+\gamma)-f_0(\varphi_2+\gamma)], \end{aligned}$$

whence:

$$\begin{aligned} \|h_1-h_2\|_2&\leqq M_fL_g\|\psi_1-\psi_2\|_2+M_gL_f\|\varphi_1-\varphi_2\|_2,\\ &\leqq[M_fL_g+M_gL_f]\|(\varphi_1-\varphi_2,\psi_1-\psi_2)\|_2, \end{aligned}$$

and the proof is complete. □

5.3. Uniqueness. The uniqueness of the solution (Θ, Y) of problem (5.1), (5.2) is a consequence of the following proposition.

PROPOSITION 5.9. *Let* $T>0$. *Under the hypotheses* (3.9), *there exists at most one solution of problem* (5.5) *in* $C([0,T), L^2\times L^2)\cap L^\infty([0,T), L^\infty\times L^\infty)$.

Proof. Let $T>0$, and let (φ_1,ψ_1) and (φ_2,ψ_2) be two solutions of (5.5), with $\varphi_i, \psi_i\in L^\infty([0,T), L^\infty)$ for $i=1,2$. Choosing $U\in\mathbb{R}$ such that $\|(\varphi_i,\psi_i)(t)\|_\infty\leqq U$ for $i=1,2$ and $t\in[0,T)$, we can consider two functions f_U and g_U satisfying:

(5.9) f_U is positive, bounded and Lipschitz-continuous on $\mathbb{R}$, $f_U(\xi)=f(\xi)$ if $|\xi|\leqq U$.

(5.10) g_U is positive, bounded and Lipschitz-continuous on $\mathbb{R}$, $g_U(\xi) = g(\xi)$ if $|\xi| \leqq U$.

(φ_1, ψ_1) and (φ_2, ψ_2) are then solutions of the following problem:

$$\begin{aligned} &\varphi_t - \varphi_{xx} = f_U(\varphi+\gamma) g_U(\psi+\gamma_1) + \gamma_{xx}, \\ &\psi_t - \frac{\psi_{xx}}{\mathrm{Le}} = -f_U(\varphi+\gamma) g_U(\psi+\gamma_1) - \frac{\gamma_{xx}}{\mathrm{Le}}, \\ &\varphi(x,0) = \varphi_0(x), \qquad \psi(x,0) = \psi_0(x). \end{aligned} \tag{5.11}$$

Applying Lemma 5.8, we get $(\varphi_1, \psi_1) = (\varphi_2, \psi_2)$, which ends the proof. □

5.4. Global existence. We show in this section the existence of a solution (Θ, Y) of problem (5.1), (5.2).

PROPOSITION 5.10. *Assume that the hypotheses* (3.9) *hold. Then there exists* $T_{\max} \in \mathbb{R}_+^* \cup \{+\infty\}$ *such that a solution* (Θ, Y) *of problem* (5.1), (5.2) *exists on* $\mathbb{R} \times [0, T_{\max})$. *Moreover,* (Θ, Y) *is a classical solution on* $(0, T_{\max})$, *and the following alternative holds*:

$$\begin{aligned} &\textit{Either}: \quad T_{\max} = +\infty, \\ &\textit{Or}: \quad \lim_{t \to T_{\max}} \|\Theta(t)\|_\infty = +\infty. \end{aligned} \tag{5.12}$$

The proof of this proposition is divided into two lemmas.

LEMMA 5.11. *Under the hypotheses* (3.9), *there exists* $T_{\max} \in \mathbb{R}_+^* \cup \{+\infty\}$ *such that a solution* (φ, ψ) *of problem* (5.5) *exists on* $\mathbb{R} \times [0, T_{\max})$. *Moreover, the following properties hold*:

$$\varphi, \psi \in C([0, T_{\max}), L^2) \cap C((0, T_{\max}), H^2) \cap C^1((0, T_{\max}), L^2), \tag{5.13}$$

$$\forall T < T_{\max}, \quad \varphi, \psi \in L^\infty([0, T), L^\infty), \tag{5.14}$$

$$\begin{aligned} &\textit{Either}: \quad T_{\max} = +\infty, \\ &\textit{Or}: \quad \lim_{t \to T_{\max}} \|(\varphi, \psi)(t)\|_\infty = +\infty. \end{aligned} \tag{5.15}$$

Proof. (a) Let us first show the existence of a solution on $\mathbb{R} \times [0, T)$ for small positive T. For $U \geqq \|(\varphi_0, \psi_0)\|_\infty + 2$, we define f_U and g_U as in (5.9), (5.10) above and consider again the problem (5.11). Lemma 5.8 applies again and gives a solution (φ_U, ψ_U). Denoting

$$F_U(\varphi_U, \psi_U) = \left[f_U(\varphi_U+\gamma) g_U(\psi_U+\gamma_1) + \gamma_{xx}, -f_U(\varphi_U+\gamma) g_U(\psi_U+\gamma_1) - \frac{\gamma_{xx}}{\mathrm{Le}} \right],$$

and using Remark 5.2, we get:

$$\begin{aligned} &(\varphi_U, \psi_U)(t) = S^2(t)(\varphi_0, \psi_0) + \int_0^t S^2(t-s) F_U[(\varphi_U, \psi_U)(s)]\, ds, \\ &\|(\varphi_U, \psi_U)(t)\|_\infty \leqq \|(\varphi_0, \psi_0)\|_\infty + \int_0^t \|F_U[(\varphi_U, \psi_U)(s)]\|_\infty\, ds. \end{aligned} \tag{5.16}$$

Since f_U, g_U and γ_{xx} are bounded, we can obviously find a constant C_U such that: for all $(\varphi_1, \psi_1) \in L^\infty \times L^\infty$, $\|F_U(\varphi_1, \psi_1)\|_\infty \leqq C_U$. This implies:

$$\|(\varphi_U, \psi_U)(t)\|_\infty \leqq \|(\varphi_0, \psi_0)\|_\infty + C_U t.$$

Let $t_U = 1/C_U$. For $t \in [0, t_U)$, we have

$$\|(\varphi_U, \psi_U)(t)\|_\infty \leqq \|(\varphi_0, \psi_0)\|_\infty + 1,$$

whence

$$\|\varphi_U(t)+\gamma\|_\infty \leqq \|(\varphi_0, \psi_0)\|_\infty + 2 \leqq U, \qquad \|\psi_U(t)+\gamma_1\|_\infty \leqq \|(\varphi_0, \psi_0)\|_\infty + 2 \leqq U.$$

This implies that (φ_U, ψ_U) is a solution of (5.5) on $\mathbb{R}\times[0, t_U)$; this solution satisfies:

$$\varphi_U, \psi_U \in C([0, t_U), L^2) \cap L^\infty([0, t_U), L^\infty),$$

and

$$\varphi_U, \psi_U \in C((0, t_U), H^2) \cap C^1((0, t_U), L^2).$$

(b) Since a solution of problem (5.5) exists locally in the neighbourhood $[0, t_U)$ of 0, it is classical to show the existence of a solution (φ, ψ) satisfying (5.13)-(5.15) on a maximal interval $[0, T_{\max})$. For sake of completeness we briefly recall the proof of (5.15): let us assume that $T_{\max} < +\infty$ and that there exists a sequence $(t_m)_{m\in\mathbb{N}}$ such that:

$$\begin{aligned} &\lim_{m\to\infty} t_m = T_{\max}, \\ &\exists V > 0, \quad \forall m \in \mathbb{N} \quad \|(\varphi, \psi)(t_m)\|_\infty \leqq V. \end{aligned} \tag{5.17}$$

Let $U = V+2$. For $m\in\mathbb{N}$, we can argue as in (a) above to show the existence of a solution of (5.5) on the interval $[t_m, t_m + t_U)$. Since t_U does not depend on m we can choose the latter so that: $t_m + t_U > T_{\max}$, which contradicts the assumption that $[0, T_{\max})$ is a maximal interval for the existence of a solution of (5.5). Formula (5.17) is therefore wrong and the alternative (5.15) holds. □

The solution (φ, ψ) of (5.5) defined in Lemma 5.11 satisfies the boundary conditions (5.6) on $(0, T_{\max})$. For $t\in(0, T_{\max})$, we have indeed $\varphi(\pm\infty, t) = \psi(\pm\infty, t) = 0$ since $\varphi, \psi \in H^1$ (see [3]).

We can now end the proof of Proposition 5.10 by using the maximum principle for parabolic partial differential equations.

LEMMA 5.12. *Let* (φ, ψ) *be the solution of* (5.5) *defined in Lemma* 5.11. *For* $(x, t)\in\mathbb{R}\times[0, T_{\max})$, *define*:

$$\Theta(x, t) = \varphi(x, t) + \gamma(x), \qquad Y(x, t) = \psi(x, t) + \gamma_1(x).$$

Then the following inequalities hold:

$$\Theta(x, t) \geqq 0, \quad 0 \leqq Y(x, t) \leqq 1 \quad \text{a.e. on } \mathbb{R}\times[0, T_{\max}). \tag{5.18}$$

Proof. (a) Let us first show that $Y \geqq 0$. This is essentially the maximum principle. For any function Z of $L^2_{\text{loc}}(\mathbb{R})$ we define as usual: $Z^- = \max(0, -Z)$, $Z^+ = \max(0, Z)$. For $t\in(0, T_{\max})$, it is known that $\psi^-(t)\in H^1$, $(\psi(t)+\gamma_1)^- \in H^1_{\text{loc}}(\mathbb{R})$ (see [12]). It follows easily from the properties (3.7) of γ and γ_1 that $(\psi(t)+\gamma_1)^- = Y^- \in H^1$. Since (φ, ψ) is a *classical solution* of (5.5) on $(0, T_{\max})$, we can write:

$$Y_t Y^- - \frac{Y_{xx} Y^-}{\text{Le}} = -f(\Theta) g(Y) Y^-.$$

But $g(Y) Y^- \equiv 0$ from (5.4); integrating the last relation by parts, we get:

$$\frac{d}{dt}\left[\frac{1}{2}\int_{\mathbb{R}} (Y^-)^2\right] + \frac{1}{\text{Le}}\int_{\mathbb{R}} [(Y^-)_x]^2 = 0,$$

whence

$$\frac{d}{dt}\left[\int_{\mathbb{R}} (Y^-)^2\right] \leqq 0 \quad \text{for } t\in(0, T_{\max}). \tag{5.19}$$

On the other hand, it can be checked easily that the mapping $\psi \to (\psi + \gamma_1)^-$ is continuous from L^2 into itself. Thus $Y^- \in C([0, T_{\max}), L^2)$. Since $\int_{\mathbb{R}} (Y^-)^2$ is decreasing on $(0, T_{\max})$ from (5.19) and $\int_{\mathbb{R}} [Y^-(t=0)]^2 = 0$ from (3.9), we obtain:

$$Y^-(t) \equiv 0 \quad \text{for } t \in [0, T_{\max}),$$

or equivalently:

$$Y(t) \geqq 0 \quad \text{for } t \in [0, T_{\max}).$$

(b) Using $(Y-1)^+$ and Θ^- instead of Y^- gives the other inequalities (5.18) as in (a) above. □

5.5. Regularity of the solution. Before showing that a global solution does exist (i.e. $T_{\max} = +\infty$), we can investigate the smoothness of the solution (Θ, Y) defined in Proposition 5.10; this is the aim of this section.

A first result concerning the regularity of the solution is the next lemma, which is an obvious consequence of Theorem 4.3 and Lemma 5.11.

LEMMA 5.13. *If* $(\varphi_0, \psi_0) \in H^2 \times H^2$, *the solution* (Θ, Y) *defined in Proposition* 5.10 *is a classical solution on* $[0, T_{\max})$.

Without any further assumption on f, we also have the following lemma.

LEMMA 5.14. *The solution* (Θ, Y) *defined in Proposition* 5.10 *satisfies*:

$$(\Theta, Y) \in C([0, T_{\max}), L^\infty). \tag{5.20}$$

Proof. Since the imbedding $H^2 \subset L^\infty$ is continuous, we already have: $(\Theta, Y) \in C((0, T_{\max}), L^\infty)$ from (5.13). Therefore we only have to show that:

$$\|(\varphi, \psi)(t) - (\varphi_0, \psi_0)\|_\infty \to 0 \quad \text{when } t \to 0. \tag{5.21}$$

We use again the notation of the proof of Lemma 5.11. Let $U > \|(\varphi_0, \psi_0)\|_\infty + 2$. For $t > 0$ small enough, (φ, ψ) is a solution of (5.11) and (5.16) implies:

$$\|(\varphi, \psi)(t) - (\varphi_0, \psi_0)\|_\infty \leqq \|S^2(\varphi_0, \psi_0) - (\varphi_0, \psi_0)\|_\infty + C_U t,$$

and formula (5.21) follows now immediately from Remark 5.2. □

The next proposition shows that, with the additional assumptions (3.15) on f, there exists a *smooth solution* of (5.1), (5.2).

PROPOSITION 5.15. *Under the hypotheses* (3.15) *on* f, *the solution* (Θ, Y) *of problem* (5.1), (5.2) *defined in Proposition* 5.10 *is a smooth solution on* $\mathbb{R} \times \mathbb{R}_+^*$. *The corresponding solution* (φ, ψ) *of* (5.5), (5.6) *satisfies*:

$$\varphi, \psi \in C((0, T_{\max}), H^4) \cap C^1((0, T_{\max}), H^2) \cap C^2((0, T_{\max}), L^2).$$

Remark 5.16. This regularity result holds without any assumption on the regularity of the initial data (φ_0, ψ_0)—only (3.9) is assumed. This is of course related to the strong regularizing effect of the heat equation.

COROLLARY 5.17. *Assume that the hypotheses* (3.15) *hold, and that* $\varphi_0, \psi_0 \in H^4$. *Then the solution* (Θ, Y) *of problem* (5.1), (5.2) *defined in Proposition* 5.10 *is a smooth solution on* $\mathbb{R} \times \mathbb{R}_+$. *The corresponding solution* (φ, ψ) *of* (5.5), (5.6) *satisfies*:

$$\varphi, \psi \in C([0, T_{\max}), H^4) \cap C^1([0, T_{\max}), H^2) \cap C^2([0, T_{\max}), L^2).$$

We begin the proof of Proposition 5.15 with two lemmas. Assuming (3.15), we first introduce two functions $\hat{f}$ and $\hat{g}$ satisfying:

$$\hat{f} \in C^2(\mathbb{R}, \mathbb{R}),$$
$$\forall \xi > 0, \quad \hat{f}_{xx} \text{ is Lipschitz-continuous on } [-\xi, \xi],$$
$$\forall \xi > 0, \quad \hat{f}(\xi) = f(\xi),$$
$$\hat{g}(\xi) = \xi^n,$$

and a mapping $\hat{F}$ defined by:

$$\hat{F}(\varphi, \psi) = \left[\hat{f}(\varphi + \gamma)\hat{g}(\psi + \gamma_1) + \gamma_{xx}, -\hat{f}(\varphi + \gamma)\hat{g}(\psi + \gamma_1) - \frac{\gamma_{xx}}{\mathrm{Le}} \right],$$

for $\varphi, \psi \in L^2$.

LEMMA 5.18. *Under the hypotheses* (3.9) *and* (3.15), *the mapping* $\hat{F}$ *is Lipschitz-continuous from any bounded subset of* $H^2 \times H^2$ *into* $H^2 \times H^2$.

Proof. (a) Let us first show that $\hat{F}(\varphi, \psi) \in H^2 \times H^2$ when $\varphi, \psi \in H^2$. For $\varphi, \psi \in H^2$, let $\hat{h} = \hat{f}(\varphi + \gamma)\hat{g}(\psi + \gamma_1)$, $M = \|(\varphi, \psi)\|_\infty$. We define: $M_f = \max_{[-M,M]} \hat{f}$, $M_g = \max_{[-M,M]} \hat{g}$, $L_f = \max_{[-M,M]} \hat{f}_x$, $L_g = \max_{[-M,M]} \hat{g}_x$. Thus $\hat{h} \in L^2$ as in the proof of Lemma 5.8. Furthermore, we have:

$$\hat{h}_x = \hat{f}_x(\varphi + \gamma)(\varphi_x + \gamma_x)\hat{g}(\psi + \gamma_1) + \hat{f}(\varphi + \gamma)\hat{g}_x(\psi + \gamma_1)(\psi_x + \gamma_{1_x}),$$
$$|\hat{h}_x| \leqq L_f M_g |\varphi_x + \gamma_x| + M_f L_g |\psi_x + \gamma_{1x}|,$$

which yields $\hat{h}_x \in L^2$. It can also be shown that $\hat{h}_{xx} \in L^2$, using the Sobolev continuous imbedding:

$$H^2 \subset W^{1,\infty}(\mathbb{R}),$$
$$\exists S > 0, \quad \forall \varphi \in H^2, \quad \|\varphi\|_{W^{1,\infty}} \leqq S \|\varphi\|_{H^2}.$$

(b) It is long but easy to check that, for any $M > 0$, $\hat{h}$ is Lipschitz-continuous from $\{(\varphi, \psi) \in H^2 \times H^2, \ \|(\varphi, \psi)\|_{H^2 \times H^2} \leqq M\}$ into H^2; the details are left to the reader. □

For $\varphi_1, \psi_1 \in L^2$, we now consider the problem:

$$\varphi_t - \varphi_{xx} = \hat{f}(\varphi + \gamma)\hat{g}(\psi + \gamma_1) + \gamma_{xx},$$
$$\psi_t - \frac{\psi_{xx}}{\mathrm{Le}} = -\hat{f}(\varphi + \gamma)\hat{g}(\psi + \gamma_1) - \frac{\gamma_{xx}}{\mathrm{Le}}, \tag{5.22}$$
$$\varphi(x, 0) = \varphi_1(x), \qquad \psi(x, 0) = \psi_1(x).$$

LEMMA 5.19. *Assume that* (3.9) *and* (3.15) *hold. Then there exist* $\hat{T}_{\max} \in \mathbb{R}_+^* \cup \{+\infty\}$ *such that a unique solution* $(\hat{\varphi}, \hat{\psi})$ *of problem* (5.22) *exists on* $\mathbb{R} \times [0, \hat{T}_{\max})$. *This solution satisfies*:

$$\hat{\varphi}, \hat{\psi} \in C((0, \hat{T}_{\max}), H^4) \cap C^1((0, \hat{T}_{\max}), H^2),$$

and the following alternative holds:

Either: $\hat{T}_{\max} = +\infty$,

Or: $\lim_{t \to \hat{T}_{\max}} \|(\hat{\varphi}, \hat{\psi})(t)\|_{H^2 \times H^2} = +\infty$.

Proof. From Lemma 5.18, it suffices to apply Theorem 4.4 with $H = H^2 \times H^2$, $D(A) = H^4 \times H^4$ and $F = \hat{F}$. □

We can now complete the proof of Proposition 5.15 and Corollary 5.17.

Proof. (a) Let (φ, ψ) be the solution of (5.5) defined in Lemma 5.11; for $\varepsilon \in (0, T_{\max})$, we set $(\varphi_1, \psi_1) = (\varphi, \psi)(t = \varepsilon) \in H^2 \times H^2$. Applying Lemma 5.19, we get a solution $(\hat{\varphi}, \hat{\psi})$ of (5.22), which is unique in $C((0, \hat{T}_{\max}), H^2 \times H^2)$. But it is straightforward to show that $(\varphi, \psi)(t+\varepsilon)$ is also a solution of (5.22) in $C([0, T_{\max} - \varepsilon), H^2 \times H^2)$. These two solutions coincide, and we get:

$$(\varphi, \psi)(t) = (\hat{\varphi}, \hat{\psi})(t-\varepsilon) \quad \text{for } t \in [\varepsilon, T_{\max}).$$

(b) Lemma 5.19 and (a) above obviously imply that the solution (φ, ψ) defined in Lemma 5.11 satisfies:

$$\varphi, \psi \in C((0, T_{\max}), H^4) \cap C^1((0, T_{\max}), H^2).$$

It suffices now to use the Sobolev continuous imbeddings $H^2 \subset C^1(\mathbb{R})$, $H^4 \subset C_3(\mathbb{R})$ to show that (φ, ψ) is a smooth solution of (5.5), (5.6) on $(0, T_{\max})$. To end the proof of Proposition 5.15, it remains to show that $\varphi, \psi \in C^2((0, T_{\max}), L^2)$, or equivalently that $\hat{F}(\varphi, \psi) \in C^1((0, T_{\max}), L^2 \times L^2)$; this is straightforward and is left to the reader. □

Before concluding this section, we state another lemma:

LEMMA 5.20. *Let (φ, ψ) be the solution of problem (5.5) defined in Lemma 5.11. Let $p \in [1, +\infty)$ and assume that $\varphi_0, \psi_0 \in L^p$. Then*

$$\forall t \in (0, T_{\max}), \qquad \varphi(t), \psi(t) \in L^p.$$

Proof. Let $T < T_{\max}$ and $t \in [0, T]$. We set $M = \max_{t \in [0,T]} \|\Theta(t)\|_\infty$ and define L_f such that:

$$\forall \xi \in [0, M], \qquad |f(\xi)| \leqq L_f \xi.$$

Denoting $\Omega = Y^n f(\Theta)$, we can write $\Omega \leqq L_f \Theta$ and $\Omega \leqq L_f M Y$ from (5.18). Then, using (3.7), we have:

$$\begin{aligned} \|\Omega\|_p^p &= \|\Omega\|_{L^p(-\infty,-1)}^p + \|\Omega\|_{L^p(-1,+1)}^p + \|\Omega\|_{L^p(1,\infty)}^p \\ &\leqq L_f^p \|\varphi\|_{L^p(-\infty,-1)}^p + L_f^p M^p \|\psi\|_{L^p(1,\infty)}^p + K \\ &\leqq K(\|\varphi\|_p^p + \|\psi\|_p^p + 1) \end{aligned}$$

(where K denotes a positive constant), whence:

$$\|\Omega\|_p \leqq K(\|\varphi\|_p + \|\psi\|_p + 1).$$

Using now (4.4) and Lemma 4.6, we have (we use the notation $\Omega(s)$ instead of $\Omega Y(., s), \Theta(., s)]$ for simplicity):

$$\varphi(t) = S(t)\varphi_0 + \int_0^t S(t-s)[\Omega(s) + \gamma_{zz}]\, ds,$$

$$\|\varphi(t)\|_p \leqq \|\varphi_0\|_p + \int_0^t (\|\Omega(s)\|_p + K)\, ds,$$

$$\|\varphi(t)\|_p \leqq K + K \int_0^t (\|\varphi(s)\|_p + \|\psi(s)\|_p + 1)\, ds.$$

Arguing in the same way for $\|\psi(t)\|_p$, we finally obtain:

$$\|\varphi(t)\|_p + \|\psi(t)\|_p \leqq K\left[1 + \int_0^t (\|\varphi(s)\|_p + \|\psi(s)\|_p)\, ds\right].$$

It suffices now to apply Gronwall's lemma and the proof is complete. □

Remark 5.21. With the hypotheses of Lemma 5.20, one could easily show that $(\varphi, \psi) \in C([0, T_{\max}), L^p)$.

5.6. Existence for all time. We now end this fifth section by showing that $T_{\max} = +\infty$.

PROPOSITION 5.22. *Under hypothesis* (3.10), *the solution* (Θ, Y) *of* (5.1), (5.2) *defined in Proposition* 5.10 *exists on* $\mathbb{R} \times \mathbb{R}_+$:

$$T_{\max} = +\infty. \tag{5.23}$$

Proof. For $p \in [1, +\infty)$ and $t \in (0, T_{\max})$, we can write, since (Θ, Y) is a *classical solution* on $(0, T_{\max})$:

$$\Theta_t \Theta^{p-1} - \Theta_{xx} \Theta^{p-1} = Y^n f(\Theta) \Theta^{p-1}.$$

Integrating by parts as in the proof of Lemma 5.12, we obtain:

$$\frac{1}{p} \frac{d}{dt} \left(\int_{\mathbb{R}} \Theta^p \right) \leqq \int_{\mathbb{R}} [Y^n f(\Theta) \Theta^{p-1}].$$

Formulae (3.10) and (5.18) now imply:

$$\frac{d}{dt} \left(\int_{\mathbb{R}} \Theta^p \right) \leqq C_f p \int_{\mathbb{R}} \Theta^p.$$

Let $t_0 \in (0, T_{\max})$. Applying Gronwall's lemma to the last inequality, we can write:

$$\int_{\mathbb{R}} \Theta(t)^p \leqq \int_{\mathbb{R}} \Theta(t_0)^p \, e^{pC_f(t-t_0)},$$

or:

$$\|\Theta(t)\| p \leqq \|\Theta(t_0)\|_p e^{C_f(t-t_0)}.$$

We can then take the limit $p \to \infty$ to get:

$$\|\Theta(t)\|_\infty \leqq \|\Theta(t_0)\|_\infty \, e^{C_f(t-t_0)},$$

which together with (5.12) implies $T_{\max} = +\infty$. □

Of course, from a physical standpoint, it can be thought that (5.23) holds even if (3.10) is not assumed, because of (5.12), since one may expect that the increase of the temperature is limited by the consumption of the reactant. Nevertheless, we have been able to prove rigorously the global existence of the solution only with the assumption (3.10), or in the following case.

LEMMA 5.23. *Assume that the hypotheses* (3.9) *hold. If moreover* $\mathrm{Le} = 1$, *then the solution* (Θ, Y) *of* (5.1), (5.2) *exists on* $\mathbb{R} \times \mathbb{R}_+$.

Proof. If $\mathrm{Le} = 1$, we can add the two equations (5.1) to get:

$$(Y + \Theta)_t - (Y + \Theta)_{xx} = 0.$$

A straightforward application of the maximum principle for parabolic partial differential equations yields:

$$\|(Y + \Theta)(t)\|_\infty \leqq \|(Y + \Theta)(0)\|_\infty,$$

and (5.23) follows again from (5.12). □

Remark 5.24. With the same hypothesis $\mathrm{Le} = 1$, it can be shown that $(Y + \Theta)$ converges towards 1 uniformly on $\mathbb{R}$ as t tends to $+\infty$:

$$\lim_{t \to +\infty} \| Y + \Theta - 1 \|_\infty = 0.$$

6. Existence and uniqueness for the hydrodynamical variables.

6.1. Statement of the problem and main results. We now want to consider the subsystem (2.11) for the hydrodynamical variables—density, velocity and pressure:

$$(\Theta+\alpha)\rho=1, \tag{6.1}$$

$$u_x=\Theta_t, \qquad u(-\infty, t)=u^0, \tag{6.2}$$

$$u_t+p_x=0, \qquad p(-\infty, t)=0. \tag{6.3}$$

Throughout this section, it will be assumed that hypotheses (3.9) and (3.10) hold. The solution (Θ, Y) of (5.1), (5.2) in $\mathbb{R}\times\mathbb{R}_+$ and the corresponding solution (φ, ψ) of (5.5), (5.6) are now considered given. We let $\Omega(y, t)=\Omega[Y(y, t), \Theta(y, t)]$.

We recall that the *weak* or *smooth solutions* of (6.1)-(6.3) are defined at the beginning of § 3 (in particular, the boundary condition (6.3b) is fulfilled in the sense of (3.6) for *weak solutions*). About problem (6.1)-(6.3), we are going to prove:

THEOREM 6.1. *Assume that the hypotheses* (3.9)-(3.13) *hold. Then there exists a unique weak solution* (ρ, u, p) *of* (6.1)-(6.3) *in* $\mathbb{R}\times\mathbb{R}_+$.

If moreover (3.14) *and* (3.15) *hold,* (ρ, u, p) *is a smooth solution on* $\mathbb{R}\times\mathbb{R}_+$.

In order to prove this result, we now solve the two problems (6.2) and (6.3) in sequence (solving (6.1) for the density ρ is an obvious task since $\Theta(x, t)+\alpha\geqq\alpha>0$ for all $(x, t)\in\mathbb{R}\times\mathbb{R}_+$).

6.2. Velocity.

PROPOSITION 6.2. *There exists a unique weak solution of* (6.2) *in* $\mathbb{R}\times\mathbb{R}_+$:

$$u(x, t)=u^0+\Theta_x(x, t)+\int_{-\infty}^{x}\Omega(y, t)\,dy \quad \textit{for } (x, t)\in\mathbb{R}\times\mathbb{R}_+ \tag{6.4}$$

and u is a smooth solution of (6.2) *in* $\mathbb{R}\times\mathbb{R}_+^*$.

Moreover, if $\varphi_0, \psi_0\in H^2$, *u is a smooth solution of* (6.2) *in* $\mathbb{R}\times\mathbb{R}_+$.

Proof. (a) Let $t>0$. Since (Θ, Y) is a *classical solution* of (5.1) in the neighbourhood of t, we have from (6.2):

$$u_x=\Theta_t=\Theta_{xx}+\Omega\in L^1_{\mathrm{loc}}(\mathbb{R}),$$

whence:

$$u(x, t)=u(0, t)+\Theta_x(x, t)-\Theta_x(0, t)+\int_0^x\Omega(y, t)\,dy.$$

Since we want a finite limit $u(-\infty, t)$ to exist, we only need to show that:

$$\int_{-\infty}^{0}\Omega(y, t)\,dy<+\infty. \tag{6.5}$$

But (3.10) and (5.18) imply: $\Omega(y, t)\leqq C_f\Theta(y, t)$ and (6.5) follows since (3.9) and Lemma 5.20 imply $\Theta(t)\in L^1(\mathbb{R}_-^*)$. We then obtain (6.4) for $t>0$ (we have $\Theta_x(-\infty, t)=0$ since $\Theta(t)\in H^2$).

(b) It is clear that the solution u defined by (6.4) for $t>0$ satisfies (6.2a) in the sense of $D'(\mathbb{R}\times\mathbb{R}_+)$. When $\varphi_0, \psi_0\in H^2$, we can argue as in (a) above for $t=0$ and obtain (6.4) for $t\geqq 0$. To show that u is then a *smooth solution* of (6.2), it remains to prove that $\int_{-\infty}^{x}\Omega(y, t)\,dy$ is continuous with respect to both variables x and t; this will be a consequence of the next lemma. □

Before studying the pressure problem (6.3), we state some results about the regularity of the velocity u.

LEMMA 6.3. *The solution u of* (6.2) *satisfies*:

$$u \in C(\mathbb{R}_+^*, C(\mathbb{R})) \cap C(\mathbb{R}_+^*, L^\infty).$$

If moreover $\varphi_0, \psi_0 \in H^2$, *then* $u \in C(\mathbb{R}_+, C(\mathbb{R}) \cap C(\mathbb{R}_+, L^\infty)$.

Proof. For $T>0$ define $M = \max_{t\in[0,T]} \|\Theta(t)\|_\infty$. For $t \in (0, T]$, we first have $u(t) \in L^\infty$, or equivalently $\Omega(t) \in L^1$ from the estimates:

$$\Omega(t) \leqq C_f \Theta(t) \in C(\mathbb{R}_+^*, L^1(\mathbb{R}_-^*)), \qquad \Omega(t) \leqq C_f M Y(t) \in C(\mathbb{R}_+^*, L^1(\mathbb{R}_+^*)).$$

These two inequalities can be written together in the form $\Omega(t) \leqq G(t)$ with $G(t) \in C(\mathbb{R}_+^*, L^1)$. The continuity of the integral $\int \Omega(y, t)\, dy$ with respect to the variable t is now a consequence of classical convergence results from integration theory. For the sake of completeness we sketch the arguments: arguing by contradiction, we assume that there exists a sequence (t_n) satisfying $t_n \to t_0 > 0$ and:

$$\|\Omega(t_n) - \Omega(t_0)\|_1 \geqq \varepsilon > 0. \tag{6.6}$$

Then from the converse of Lebesgue's bounded convergence theorem (see [3, p. 58]), there exists $G_0 \in L^1$ and a subsequence (t_{n_k}) such that $G(y, t_{n_k}) \leqq G_0(y)$ a.e. for all n_k. Since (5.20) proves that $\Omega(t_{n_k})$ converges pointwise towards $\Omega(t_0)$, Lebesgue's bounded convergence theorem now shows that $\Omega(t_{n_k})$ converges to $\Omega(t_0)$ in L^1, which contradicts (6.6) and ends the proof. □

LEMMA 6.4. *Assume that the hypotheses* (3.12) *hold and define* $v(x, t) = \int_{-\infty}^{x} \Omega(y, t)\, dy$. *Then*:

$$v \in C^1(\mathbb{R} \times \mathbb{R}_+^*, \mathbb{R}) \quad \text{and} \quad v_t(x, t) = \int_{-\infty}^{x} \Omega_t(y, t)\, dy.$$

Proof. Assumption (3.12) obviously implies:

$$\forall M > 0, \quad \exists K_M > 0, \quad \forall \vartheta \in [0, M] \quad |f'(\vartheta)| \leqq K_M \vartheta^\beta \tag{6.7}$$

(with $\beta > \frac{1}{2}$). Let again $T>0$ and $M = \max_{t\in[0,T]} \|\Theta(t)\|_\infty$. For $t \in [0, T]$ and $y \in \mathbb{R}$ we have:

$$\Omega_t = nY^{n-1} Y_t f(\Theta) + Y^n \Theta_t f'(\Theta),$$

whence:

$$|\Omega_t(t)| \leqq nC_f \Theta(t) |Y_t(t)| + K_M \Theta^\beta(t) |\Theta_t(t)|,$$

$$|\Omega_t(t)| \leqq K[\Theta^2 + Y_t^2 + \Theta^{2\beta} + \Theta_t^2](t) \in C(\mathbb{R}_+^*, L^1(\mathbb{R}_-^*)),$$

where K is a positive constant. The proof is then completed in a way similar to that of the previous lemma. □

The next result is now an obvious consequence of the above lemmas and of Proposition 5.15.

LEMMA 6.5. *Under hypotheses* (3.12) *and* (3.15), *the solution u of* (6.2) *satisfies*:

$$u \in C(\mathbb{R}_+^*, C^2(\mathbb{R})) \cap C^1(\mathbb{R}_+^*, C(\mathbb{R})).$$

If moreover $\varphi_0, \psi_0 \in H^4$, *then* $u \in C(\mathbb{R}_+, C^2(\mathbb{R})) \cap C^1(\mathbb{R}_+, C(\mathbb{R}))$.

6.3. Pressure. We now investigate the problem (6.3) for the pressure. We are going to prove the following.

PROPOSITION 6.6. *Assume that the hypotheses* (3.11)-(3.13) *hold and that* $\varphi_0, \psi_0 \in H^2$. *Then there exists a unique weak solution of* (6.3) *on* $\mathbb{R}\times\mathbb{R}_+$.

If moreover the hypotheses (3.14) *and* (3.15) *hold, there exists a unique smooth solution of* (6.3) *on* $\mathbb{R}\times\mathbb{R}_+$.

Proof. (a) If p is a solution of (6.3), we get from (6.4) and Lemma 6.4:

$$p_x = -\Theta_{xt} - \int_{-\infty}^{x} \Omega_t,$$

whence:

$$p = -\Theta_t - \int_{-\infty}^{x} \left[dy \int_{-\infty}^{y} \Omega_t \right],$$

or

$$p(x,t) = -\Theta_{xx}(x,t) - \Omega(x,t) - \int_{-\infty}^{x} \left[dy \int_{-\infty}^{y} \Omega_t(z,t)\, dz \right] \tag{6.8}$$

because of the boundary condition (6.3b).

Therefore, we need to prove that the last integral does exist when the assumptions (3.11)-(3.13) hold. This amounts to showing that $\Omega_t(y,t)$ vanishes at $-\infty$ at least as fast as some negative power of y. More precisely, we are going to show that:

$$\exists \varepsilon > 0, \quad \forall y < 0, \quad \int_{-\infty}^{y} |\Omega_t| \leqq \frac{1}{|y|^{1+\varepsilon}}. \tag{6.9}$$

We first need to introduce the functional space:

$$W_\nu = \{ w \in L^2 \cap L^\infty, \max_{y \in \mathbb{R}^*_-} |y^\nu w(y)| < +\infty \}$$

for $\nu > 0$, with the norm: $\|w\|_{W_\nu} = \|w\|_2 + \|w\|_\infty + \|y^\nu w\|_{L^\infty(\mathbb{R}^*_-)}$, and to state the next lemma, which is proved at the end of this section.

LEMMA 6.7. *Let* $\nu > 0$ *be given. If* $\varphi_0 \in W_\nu$, *then* $\varphi(t) \in W_\nu$ *for any* $t > 0$.

We now have:

$$|\Omega_t| = |Y^n f'(\Theta)\Theta_t + nY^{n-1} Y_t f(\Theta)| \leqq |f'(\Theta)||\Theta_t| + n|Y_t||f(\Theta)|.$$

Let $T > 0$ and $M = \max_{t\in[0,T]} \|\Theta(t)\|_\infty$. Since $(\Theta_t, Y_t) \in C([0,T], L^2 \times L^2)$, we can set $M' = \max_{t\in[0,T]} \|\Theta_t(t)\|_2$. For $t \in [0,T]$ and $y < 0$ we have:

$$\int_{-\infty}^{y} |f'(\Theta)||\Theta_t| \leqq \left[\int_{-\infty}^{y} f'(\Theta)^2 \right]^{1/2} \left[\int_{-\infty}^{y} \Theta_t^2 \right]^{1/2}$$

by the Cauchy-Schwarz inequality. Hence, using (6.7):

$$\int_{-\infty}^{y} |f'(\Theta)||\Theta_t| \leqq M' K_M \left[\int_{-\infty}^{y} \Theta^{2\beta} \right]^{1/2}.$$

As $\varphi_0 \in W_\mu$ with $\beta\mu > 3/2$ from (3.13), we can apply Lemma 6.7 to get:

$$\int_{-\infty}^{y} |f'(\Theta)||\Theta_t| \leqq K \left[\int_{-\infty}^{y} \frac{dz}{|z|^{2\beta\mu}} \right]^{1/2} \leqq \frac{K}{|y|^{\beta\mu - 1/2}}.$$

Since (6.7) implies $f(\Theta) \leqq K'_M \Theta^{\beta+1}$ we can argue in the same way for $\int_{-\infty}^{y} |f(\Theta)||Y_t|$ and (6.9) holds.

(b) It is straightforward to check that p defined by (6.8) is a solution of (6.3a) in the sense of $D'(\mathbb{R}\times\mathbb{R}_+)$ (it suffices to argue as in Remark 5.5 and to use Lemma 6.3 for the continuity of u in the neighbourhood of $t=0$). (u, p) is the unique *weak solution* of (6.3) in the sense of Definition 3.1. Furthermore, if (3.14) and (3.15) hold, (6.3a) is fulfilled in the classical sense and p is a *smooth solution* of (6.3). □

It remains now to prove Lemma 6.7. We begin the proof with a property of the linear semigroup $S(t)$ generated by the heat operator (see the end of § 4).

LEMMA 6.8. *Let $\nu>0$. The operator $S(t)$ maps W_ν into itself: for any $T>0$, there exists a positive constant M_T such that:*

$$\forall w_0\in W_\nu,\quad \forall t\in[0,T],\quad \|S(t)w_0\|_{W_\nu}\leqq M_T\|w_0\|_{W_\nu}.$$

Proof. Let $\nu>0$, $w_0\in W_\nu$, $T>0$, $t\in[0,T]$. Lemma 4.6 implies:

$$\|S(t)w_0\|_2+\|S(t)w_0\|_\infty\leqq\|w_0\|_2+\|w_0\|_\infty.$$

Therefore it remains to study $\|y^\nu S(t)w_0\|_{L^\infty(\mathbb{R}^*_\pm)}$.

Let $x<0$; we have from (4.5):

$$|x|^\nu S(t)w_0(x)=\frac{|x|^\nu}{\sqrt{4\pi t}}\int_{-\infty}^{x/2}w_0(y)\,e^{-|x-y|^2/4t}\,dy+\frac{|x|^\nu}{\sqrt{4\pi t}}\int_{x/2}^{+\infty}w_0(y)\,e^{-|x-y|^2/4t}\,dy.$$

Let us denote by $A(x)$ and $B(x)$ the two terms in the right-hand side of this relation. For $y\in(-\infty, x/2]$, we have:

$$|w_0(y)|\leqq\frac{\|w_0\|_{W_\nu}}{|y|^\nu}\leqq 2^\nu\frac{\|w_0\|_{W_\nu}}{|x|^\nu}.$$

Thus:

$$|A(x)|\leqq\frac{2^\nu}{\sqrt{4\pi t}}\|w_0\|_{W_\nu}\int_{\mathbb{R}}e^{-|x-y|^2/4t}\,dy=2^\nu\|w_0\|_{W_\nu}.$$

On the other hand, we also have:

$$|B(x)|\leqq\frac{|x|^\nu}{\sqrt{4\pi t}}\|w_0\|_\infty\int_{x/2}^{+\infty}e^{-|x-y|^2/4t}\,dy.$$

Setting $z=(y-x)/\sqrt{4t}$ and assuming $x<-4\sqrt{T}$, we obtain:

$$\begin{aligned}|B(x)|&\leqq\frac{|x|^\nu}{\sqrt{\pi}}\|w_0\|_\infty\int_{|x|/4\sqrt{t}}^{+\infty}e^{-z^2}\,dz\\&\leqq\frac{|x|^\nu}{\sqrt{\pi}}\|w_0\|_\infty\int_{|x|/4\sqrt{T}}^{+\infty}e^{-z^2}\,dz\leqq\frac{|x|^\nu}{\sqrt{\pi}}\|w_0\|_\infty\int_{|x|/4\sqrt{T}}^{+\infty}e^{-z}\,dz\\&\leqq\frac{|x|^\nu}{\sqrt{\pi}}e^{-|x|/4\sqrt{T}}\|w_0\|_\infty\leqq K\|w_0\|_\infty,\end{aligned}$$

and the proof is easily achieved. □

Proof of Lemma 6.7. Let $T>0$ and $M=\max_{t\in[0,T]}\|\Theta(t)\|_\infty$. For $t\in[0,T]$ we can write:

$$\begin{aligned}\|\Omega\|_\infty&\leqq C_f\|\Theta\|_\infty\leqq C_fM,\\\|\Omega(t)\|_2^2&\leqq\|\Omega(t)\|^2_{L^2(-\infty,-1)}+\|\Omega(t)\|^2_{L^2(-1,+1)}+\|\Omega(t)\|^2_{L^2(1,\infty)}\\&\leqq C_f^2\|\varphi(t)\|_2^2+2C_f^2M^2+C_f^2M^2\|\psi(t)\|_2\leqq K,\\\|y^\nu\Omega(t)\|_{L^\infty(\mathbb{R}^*_\pm)}&\leqq C_f\|y^\nu\Theta(t)\|_{L^\infty(\mathbb{R}^*_\pm)},\end{aligned}$$

whence: $\|\Omega(t)\|_{W_\nu}\leqq K(\|\varphi(t)\|_{W_\nu}+1)$, where K is a positive constant.

From (4.4) and (5.5), we have:

$$\varphi(t) = S(t)\varphi_0 + \int_0^t S(t-s)[\Omega(s) + \gamma_{xx}]\, ds.$$

Applying Lemma 6.8 yields:

$$\|\varphi(t)\|_{W_\nu} \leqq M_T \|\varphi_0\|_{W_\nu} + M_T \int_0^t \|\Omega(s) + \gamma_{xx}\|_{W_\nu}\, ds$$

$$\leqq K\left[1 + \int_0^t \|\varphi(s)\|_{W_\nu}\, ds\right].$$

It suffices now to apply Gronwall's lemma and the proof is complete. □

7. Back transformation to the Eulerian variables. We now want to show that the results of the preceding sections make it possible to show the existence of a solution for the Eulerian system (3.4), (3.5). Since the equivalence between *smooth solutions* of the two systems (3.1)–(3.3) and (3.4), (3.5) follows immediately from § 2, we only have to investigate the existence and uniqueness of a *weak solution* of (3.4), (3.5).

7.1. Coordinate transformation. We first need to study the change of variables between the Lagrangian and the Eulerian system. This is the aim of the next two lemmas.

LEMMA 7.1. *Assume that the hypotheses* (3.11), (3.13) *hold and let* (Θ, Y, ρ, u) *be the unique weak solution of* (3.1), (3.2). *Consider the mapping*:

$$T_{LE}: \left| \begin{array}{l} \mathbb{R}\times\mathbb{R}_+ \to \mathbb{R}\times\mathbb{R}_+, \\ (x, t) \to (\xi, \tau) \text{ defined as}: \end{array} \right.$$

$$\xi(x, t) = \int_0^t u(0, t')\, dt' + \int_0^x \frac{1}{\rho(x', t)}\, dx', \qquad \tau(x, t) = t. \tag{7.1}$$

T_{LE} is a bijection from $\mathbb{R}\times\mathbb{R}_+$ into itself. Furthermore, $T_{LE} \in C^1(\mathbb{R}\times\mathbb{R}_+^*, \mathbb{R}\times\mathbb{R}_+^*)$ and:

$$\xi_x(x, t) = \frac{1}{\rho(x, t)}, \qquad \xi_t(x, t) = u(x, t). \tag{7.2}$$

Proof. Since $u(0, t)$ and $1/\rho(x, t)$ are continuous functions on $\mathbb{R}\times\mathbb{R}_+$, the relations (7.1) define a mapping from $\mathbb{R}\times\mathbb{R}_+$ into itself. Moreover, we have $\xi_x = 1/\rho$ and, for $t > 0$:

$$\xi(x, t) = \int_0^t u(0, t')\, dt' + \int_0^x [\Theta(x', t) + \alpha]\, dx'.$$

It is clearly possible to differentiate under the second integral sign in this expression to get:

$$\xi_t(x, t) = u(0, t) + \int_0^x \Theta_t(x', t)\, dx' = u(0, t) + \int_0^x u_x(x', t)\, dx' = u(x, t).$$

On the other hand, we can easily define $T_{EL} = T_{LE}^{-1}$ by setting $T_{EL}(\xi, \tau) = (x, t)$ with:

$$x(\xi, \tau) = \int_{\xi_0(\tau)}^{\xi} \rho(\xi', \tau)\, d\xi', \qquad t(\xi, \tau) = t \tag{7.3}$$

where $\xi_0(\tau) = \int_0^\tau u(0, \tau')\, d\tau' [= \xi(0, \tau)]$. The end of the proof is now obvious and is omitted. □

We can use this lemma to define the following transformation: for any $\eta \in L^\infty_{\text{loc}}(\mathbb{R}\times\mathbb{R}_+)$ we define $\hat\eta \in L^\infty_{\text{loc}}(\mathbb{R}\times\mathbb{R}_+)$ by:

$$\hat\eta(\xi,\tau) = \eta[T_{EL}(\xi,\tau)] = \eta[x(\xi,\tau),\tau]. \tag{7.4}$$

We can then state the following.

LEMMA 7.2. *Assume that the hypotheses* (3.11)-(3.13) *hold and consider the transformation* $\eta \to \hat\eta$ *defined by* (7.4). *The following properties hold for* $p \in [1,+\infty]$ *and* $t_0 \in \mathbb{R}_+$:

$$\text{If } \eta \in C(\mathbb{R}_+, L^p) \quad \text{then } \hat\eta \in C(\mathbb{R}_+, L^p),$$
$$\text{If } \eta \in C(\mathbb{R}_+, C(\mathbb{R})) \quad \text{then } \hat\eta \in C(\mathbb{R}_+, C(\mathbb{R})),$$
$$\text{If } \eta(\pm\infty, t_0) = \eta_0 \quad \text{then } \hat\eta(\pm\infty, t_0) = \eta_0.$$

Moreover, similar properties hold for the derivatives:

$$\text{If } \eta \in C(\mathbb{R}_+, H^2) \quad \text{then } \hat\eta \in C(\mathbb{R}_+, H^2),$$
$$\text{If } \eta \in C(\mathbb{R}_+, C^2(\mathbb{R})) \quad \text{then } \hat\eta \in C(\mathbb{R}_+, C^2(\mathbb{R})),$$
$$\text{If } \eta \in C^1(\mathbb{R}_+, C(\mathbb{R})) \quad \text{then } \hat\eta \in C^1(\mathbb{R}_+, C(\mathbb{R})).$$

Proof. These properties are easy to check and their proofs are omitted. We simply indicate the expressions of the partial derivatives of η and $\hat\eta$ which will be useful in the sequel; (7.2) and (7.4) obviously imply:

$$\hat\eta_\xi = \rho\eta_x, \qquad \hat\eta_\tau = -\rho u\eta_x + \eta_t,$$
$$\eta_x = \frac{1}{\hat\rho}\hat\eta_\xi, \qquad \eta_t = \hat\eta_\tau + \hat u\hat\eta_\xi. \qquad \square$$

7.2. Equivalence between the Lagrangian and Eulerian formulations. We can now show the existence of a weak solution to the Eulerian system (3.4), (3.5).

PROPOSITION 7.3. *Assume that the assumptions* (3.11)-(3.13) *hold. Let* (Θ, Y, ρ, u, p) *be the unique weak solution of* (3.1)-(3.3) *and define* $(\hat\Theta, \hat Y, \hat\rho, \hat u, \hat p)$ *using* (7.2). *Then* $(\hat\Theta, \hat Y, \hat\rho, \hat u, \hat p)$ *is a weak solution of* (3.4), (3.5).

Proof. We only sketch the proof by studying the temperature equation. The weak solution satisfies (see Remark 5.5):

$$\int_{\mathbb{R}\times\mathbb{R}_+} [-\Theta\eta_t + \Theta_x\eta_x - \Omega\eta] = \int_{\mathbb{R}} \Theta(\cdot,0)\eta(\cdot,0), \tag{7.5}$$

for any $\eta \in D(\mathbb{R}\times\mathbb{R}_+)$. This relation also holds for $\eta \in D^1(\mathbb{R}\times\mathbb{R}_+)[\eta \in C^1(\mathbb{R}\times\mathbb{R}_+)$ with compact support], since $D(\mathbb{R}\times\mathbb{R}_+)$ is dense in $D^1(\mathbb{R}\times\mathbb{R}_+)$.

Let $\hat\eta \in D(\mathbb{R}\times\mathbb{R}_+)$ and let η be the unique function such that $\eta(x,t) = \hat\eta[T_{LE}(x,t)]$. Since $\eta \in D^1(\mathbb{R}\times\mathbb{R}_+)$, (7.5) holds. Using the change of coordinates (7.1) in (7.5) gives:

$$\int_{\mathbb{R}\times\mathbb{R}_+} \left[-\hat\rho\hat\Theta(\hat\eta_\tau + \hat u\hat\eta_\xi) + \hat\rho\frac{\hat\Theta_\xi}{\hat\rho}\frac{\hat\eta_\xi}{\hat\rho} - \hat\rho\hat\Omega\hat\eta \right] = \int_{\mathbb{R}} \hat\rho(\cdot,0)\hat\Theta(\cdot,0)\hat\eta(\cdot,0),$$

where we have used the Jacobian $\partial(x,t)/\partial(\xi,\tau) = \rho$. The last relation, which is true for any $\hat\eta \in D(\mathbb{R}\times\mathbb{R}_+)$, says that:

$$(\hat\rho\hat\Theta)_\tau + (\hat\rho\hat u\hat\Theta)_\xi - \left(\frac{\hat\Theta_\xi}{\hat\rho}\right)_\xi = \hat\rho\hat\Omega,$$

in the sense of the distributions in $\mathbb{R}\times\mathbb{R}_+$. $\square$

To end the proof of Theorem 3.6, we still need the following lemma.

LEMMA 7.4. *There exists at most one weak solution* $(\hat\Theta, \hat Y, \hat\rho, \hat u, \hat p)$ *of* (3.4), (3.5) *satisfying*:

$$\hat u \in C(\mathbb{R}\times\mathbb{R}_+, \mathbb{R}), \qquad \hat\rho \in C(\mathbb{R}\times\mathbb{R}_+, \mathbb{R}). \tag{7.6}$$

Proof. Let $(\hat{\Theta}, \hat{Y}, \hat{\rho}, \hat{u}, \hat{p})$ be a *weak solution* of (3.4), (3.5). Thanks to (7.6), the transformations (7.3) and (7.4) can be used to show (exactly as in the proof of Proposition 7.3) that $(\hat{\Theta}, \hat{Y}, \hat{\rho}, \hat{u}, \hat{p})$ corresponds to a *weak solution* (Θ, Y, ρ, u, p) of (3.1)-(3.3); the uniqueness then follows from §§ 5 and 6. □

Remark 7.5. The uniqueness of a *weak solution* of (3.4), (3.5) can also be proven without (7.6). In this case, the equivalence between Lagrangian locally bounded *weak solutions* and Eulerian locally bounded *weak solutions* still holds, but is less simple to prove (see [13]).

8. Extension to chemically complex flames. In this section we extend our analysis to the equations of a chemically complex flame propagating in a dilute premixed gaseous mixture.

8.1. Physical assumptions. We will assume that the mixture is made up of N components $A_1, A_2 \cdots A_N$, whose mass fractions are respectively $Y_1, Y_2 \cdots Y_N$. The last species A_N is chemically inert and the reactants and products are highly diluted in a bath of A_N:

$$\sum_{s=1}^{N-1} Y_s \ll Y_N. \tag{8.1}$$

It therefore makes sense to consider that the specific heat c_p and the thermal conductivity λ of the mixture are those of the inert. Also assuming that the matrix of the diffusion coefficients is diagonal (the diffusion flux for the sth component only depends on ∇Y_s), we obtain that all the species have equal diffusivities (see [5, p. 8]), a fairly classical assumption.

Let M be the number of irreversible chemical reactions taking place in the mixture. From $1 \leqq r \leqq M$, the rth reaction can be written as:

$$\sum_{s=1}^{N-1} \nu_{rs} A_s \rightarrow \sum_{s=1}^{N-1} \mu_{rs} A_s,$$

where the stoicheometric coefficients ν_{rs} and μ_{rs} are positive integers (ν_{rs} (resp., μ_{rs}) is equal to zero if the species A_s is not a reactant (resp., a product) in the rth reaction). Let ω_r be the rate at which this reaction proceeds (a relation analogous to (2.2) gives ω_r as a function of the temperature and the mass fractions Y_s).

We can now write the governing equations of the propagation of this chemically complex flame under the form:

$$\begin{aligned}
&\rho_\tau + (\rho u)_\xi = 0,\\
&\rho u_\tau + \rho u u_\xi = -p_\xi,\\
&\rho c_p T_\tau + \rho u c_p T_\xi - (\lambda T_\xi)_\xi = \sum_{r=1}^{M} Q_r \omega_r,\\
&\rho (Y_s)_\tau + \rho u (Y_s)_\xi - (\rho D (Y_s)_\xi)_\xi = m_s \sum_{r=1}^{M} (\mu_{rs} - \nu_{rs}) \omega_r \quad \text{for } 1 \leqq s \leqq N,\\
&\sum_{s=1}^{N} Y_s = 1,
\end{aligned} \tag{8.2}$$

$$\rho T = m_N \frac{P}{R}. \tag{8.3}$$

We have defined $\nu_{rN} = \mu_{rN} = 0$ for all r. The heat released by the rth reaction, which is no more assumed to be exothermic, is denoted by Q_r, and m_s is the molecular mass of the sth species; the other notation are defined as in § 2.

Remark 8.1. The form $\rho T = m_N(P/R)$ of the equation of state follows from the assumption (8.1). The perfect gas law gives the value $P_s = RT(\rho Y_s/m_s)$ for the partial pressure of each species. Using Dalton's law we get $P = \rho RT \sum_{s=1}^{N} (Y_s/m_s)$ for the total pressure P, and this last expression reduces to (8.3) in view of (8.1).

Assuming again that the Lewis number $\mathrm{Le} = \lambda/\rho c_p D$, the specific heat c_p and the ratio λ/T are constant, we can write a Eulerian and a Lagrangian normalized form of (8.2), (8.3) as follows:

$$
(8.4)\quad \begin{cases}
\rho_\tau + (\rho u)_\xi = 0, \\
(\rho u)_\tau + (\rho u^2)_\xi = -p_\xi, \\
(\rho\Theta)_\tau + (\rho u\Theta)_\xi - \left(\dfrac{\Theta_\xi}{\rho}\right)_\xi = \displaystyle\sum_{r=1}^{M} Q_r \rho \Omega_r, \\
(\rho Y_s)_\tau + (\rho u Y_s)_\xi - \dfrac{1}{\mathrm{Le}}\left[\dfrac{(Y_s)_\xi}{\rho}\right]_\xi = m_s \displaystyle\sum_{r=1}^{M} (\mu_{rs} - \nu_{rs})\rho\Omega_r \quad \text{for } 1 \leqq s \leqq N, \\
\displaystyle\sum_{s=1}^{N} Y_s = 1, \\
(\Theta + \alpha)\rho = 1.
\end{cases}
$$

$$
(8.5)\quad \begin{cases}
\Theta_t - \Theta_{xx} = \displaystyle\sum_{r=1}^{M} Q_r \Omega_r, \\
(Y_s)_t - \dfrac{(Y_s)_{xx}}{\mathrm{Le}} = m_s \displaystyle\sum_{r=1}^{M} (\mu_{rs} - \nu_{rs})\Omega_r \quad \text{for } 1 \leqq s \leqq N, \\
\displaystyle\sum_{s=1}^{N} Y_s = 1, \\
(\Theta + \alpha)\rho = 1, \\
u_x = \Theta_t, \\
u_t + p_x = 0.
\end{cases}
$$

The following boundary conditions are associated with the above systems:

$$
\begin{aligned}
&\Theta(-\infty) = 0, && \Theta(+\infty) = 1, \\
&Y_s(-\infty) = Y_{su}, && Y_s(+\infty) = Y_{sb}, \\
&u(-\infty) = u^0, && p(-\infty) = 0.
\end{aligned}
$$

Let us denote $\Omega_r = \prod_{s=1}^{N-1} Y_s^{\nu_{rs}} f_r(\Theta)$. Since we use time-independent boundary conditions, we have to assume that the two thermochemical states prescribed at the boundaries $-\infty$ and $+\infty$ correspond to equilibria, i.e.:

$$
\forall r \in \{1, 2 \cdots M\}, \quad f_r(0) = 0,
$$

$$
\forall r \in \{1, 2 \cdots M\}, \quad \prod_{s=1}^{N-1} Y_{sb}^{\nu_{rs}} = 0.
$$

It is then straightforward to extend to systems (8.4) and (8.5) the results stated in § 3 (with a change of unknowns similar to (5.3) and assumptions analogous to (3.9)-(3.15)). Stating in detail the hypotheses and the theorems would be too long, but there appears to be no new difficulty in applying the arguments of §§ 5-7 to systems (8.4) and (8.5).

Remark 8.2. The global existence and uniqueness results stated in § 3 are also easily extended to the case of the nonadiabatic propagation of a planar flame. In this case, the energy balance equation (2.3b) becomes:

$$\rho c_p T_\tau + \rho u c_p T_\xi - (\lambda T_\xi)_\xi = Q\omega(Y, T) - \kappa(T),$$

where $\kappa(T) \geqq 0$ represents the heat losses (see [5]; for instance $\kappa(T) \equiv k(T - T_{\text{ref}})$ if only conductive heat losses are considered). In Lagrangian coordinates, the energy equation (2.10a) reads as:

$$\Theta_t = \Theta_{xx} + \Omega(Y, \Theta) - \hat{\kappa}(\Theta),$$

with $\hat{\kappa}(\Theta) \geqq 0$, $\hat{\kappa}(0) = 0$, and the arguments presented in §§ 5 and 6 apply.

REFERENCES

[1] M. T. AIMAR, *Etude numérique d'une équation d'évolution non linéaire décrivant l'instabilité thermo-diffusive d'un front de flamme*, thesis, Univ. de Provence, Marseille, 1983.

[2] H. BERESTYCKI, B. NICOLAENKO AND B. SCHEURER, *Traveling wave solutions to combustion models and their singular limits*, this Journal, 16 (1985), pp. 1207-1242.

[3] H. BREZIS, *Analyse fonctionnelle: théorie et applications*, Masson, Paris, 1982.

[4] ———, *Equations d'évolution non linéaires*, to appear.

[5] J. D. BUCKMASTER AND G. S. S. LUDFORD, *Theory of Laminar Flames*, Cambridge Univ. Press, Cambridge, England, 1982.

[6] P. EMBID, *Well-posedness of the nonlinear equations for zero mach number combustion*, Ph.D. thesis, Univ. of California, Berkeley, CA, 1984.

[7] G. S. S. LUDFORD, *Combustion: basic equations and peculiar asymptotics*, J. Méc., 16 (1977), pp. 531-551.

[8] M. MARION, *Sur les équations de flamme laminaire sans température d'ignition*, thesis, Univ. de Paris VI, Paris, 1983.

[9] B. NICOLAENKO, B. SCHEURER AND R. TEMAM, *Some global dynamical properties of the Kuramoto-Sivashinsky equations: nonlinear stability and attractors*, Phys. D, 16 (1985), pp. 155-183.

[10] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[11] N. PETERS, *Discussion of Test Problem* A, *Numerical Methods in Laminar Flame Propagation*, N. Peters and J. Warnatz, eds., Vieweg, Wiesbaden, 1982, pp. 1-14.

[12] G. STAMPACCHIA, *Equations elliptiques du second ordre à coefficients discontinus*, Presses Univ. Montreal, Montreal, 1966.

[13] D. H. WAGNER, *Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions*, to appear.

[14] F. A. WILLIAMS, *Combustion Theory*, 2nd ed., Benjamin-Cummings, Menlo Park, CA, 1985.

[15] K. YOSIDA, *Functional Analysis*, 2nd ed., Springer-Verlag, New York-Heidelberg, 1968.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

004

THE ACOUSTIC APPROXIMATION FOR COMPRESSIBLE FLOW IN THE PRESENCE OF A SURFACE UNDERGOING SMALL AMPLITUDE VIBRATIONS*

JEFFERY COOPER†

Abstract. Existence and uniqueness for short time is proved for the solutions of compressible isentropic flow in a bounded region with a moving boundary. These solutions have an asymptotic expansion in η, the amplitude of the boundary motion, as $\eta \to 0$. The leading term in this expansion is a constant flow in a fixed region, and the second term is a solution of the linear acoustic equations in the fixed region which satisfies an inhomogeneous boundary condition.

Key words. acoustics, moving boundary

AMS(MOS) subject classification. 35

1. Introduction. In this paper we shall make a rigorous derivation of the equation and boundary conditions which describe the acoustic waves produced and reflected by a vibrating surface when the amplitude of the vibrations is small. The sound waves produced by a loudspeaker are a typical example of this situation.

In most engineering textbooks, e.g., [3, p. 100], the derivation begins with the linear wave equation for the velocity potential, $\phi(x, t)$, $x \in \mathbb{R}^3$, $t \in \mathbb{R}$. Denote the moving 2-dimensional surface by $S(t)$ and assume that $S(t) = S_0 + r(t)$ where S_0 is a fixed surface and $r(t)$ is a displacement vector. Then the usual acoustic boundary condition would be

$$\frac{\partial \phi}{\partial n} = n \cdot \partial_t r \quad \text{on } S(t). \tag{1.1}$$

This boundary condition, however, leads to an ill-posed problem for the linear wave equation [1]. Acoustic engineers have found that the "correct" boundary condition for the linear wave equation is obtained by assuming that $S(t)$ oscillates about S_0 and that the wavelength of sound is much longer than the amplitude of the oscillations of $S(t)$. One then requires that

$$\frac{\partial \phi}{\partial n} = n \cdot \partial_t r \quad \text{on } S_0. \tag{1.2}$$

Recent work by Majda [2] and Schochet [5] for the nonlinear equations of compressible flow allow us to make a more systematic derivation of this boundary condition. In § 2 we prove the short-time existence of solutions of the equations of an isentropic, compressible gas in three space dimensions in a bounded region with a moving boundary. The boundary condition is that particles of the gas do not cross the moving boundary surface.

In § 3 we introduce a small parameter η which is a measure of the amplitude of the boundary motion. We then rigorously prove in Theorem 2 that the solution U_η of the nonlinear equations has an asymptotic expansion

$$U_\eta = U_0 + \eta U_1 + o(\eta) \quad \text{as } \eta \to 0$$

* Received by the editors September 2, 1986; accepted for publication (in revised form) January 23, 1987.
† Department of Mathematics, University of Maryland, College Park, Maryland 20742.

where U_0 is the state of zero velocity and constant density, and U_1 is the solution of the linearized acoustic equation with the boundary condition

$$n \cdot U_1 = n \cdot \partial_t r \quad \text{on } S_0.$$

When a velocity potential ϕ is introduced, this is precisely (1.2).

2. Nonlinear equations. For each $t \geqq 0$, let $\Omega(t)$ be a smoothly bounded open set of $\mathbb{R}^3$ with $\bar{\Omega}(t)$ compact. We set

$$Q = \bigcup_{t>0} \Omega(t) \times \{t\} \quad \text{and} \quad \Sigma = \bigcup_{t \geqq 0} \partial\Omega(t) \times \{t\}.$$

Set $Q_T = Q \cap \{0 < t < T\}$ and $\Sigma_T = \Sigma \cap \{0 \leqq t \leqq T\}$. The smoothness of the motion of the $\Omega(t)$ will come from assuming that there is a fixed, bounded, open set $\tilde{\Omega} \subset \mathbb{R}^3$, with $\partial\tilde{\Omega}$ smooth and a family of smooth mappings $x \to \phi(x, t) : \Omega(t) \to \tilde{\Omega}$, with inverse $y \to \psi(y, t) : \tilde{\Omega} \to \Omega(t)$, such that $(y, t) \to (\psi(y, t), t)$ extends to a diffeomorphism of an open neighborhood of $\tilde{\Omega} \times [0, \infty)$ onto an open neighborhood of $\bar{Q}$. $v = \partial_t \psi(\phi(x, t), t)$ is the velocity at a point $(x, t) \in \bar{Q}$, and we assume that

$$\sup_{\tilde{\Omega} \times [0, \infty)} |\partial_t \psi| < \infty. \tag{2.1}$$

Finally, we let $\nu = (\nu_x, \nu_t)$ be the space-time unit normal to Σ.

For a function $U : Q \to \mathbb{R}^p$, we shall let

$$\tilde{U}(y, t) = U(\psi(y, t), t)$$

when it is necessary to indicate that we are considering U in the coordinate system of $\tilde{Q}$. When no confusion will arise, we suppress the symbol $\tilde{}$ to simplify the notation.

We shall need the following spaces. $H^s(\tilde{\Omega})$, $s \geqq 0$, will denote the usual Sobolev space. Let $X_{m,T}$ be the space of functions U on Q such that

$$\tilde{U} \in \bigcap_{j=0}^{m} C^j([0, T]; H^{m-j}(\tilde{\Omega})).$$

We set

$$|||U(t)|||_m^2 = |||\tilde{U}(t)|||_m^2 = \sum_{j=0}^{m} \|\partial_t^j \tilde{U}(t)\|_{H^{m-j}(\tilde{\Omega})}^2$$

and give $X_{m,T}$ the norm

$$|||U|||_{m,T} = \sup_{0 \leqq t \leqq T} |||U(t)|||_m.$$

Finally, we let $Y_{\delta,T}$ $(0 < \delta < 1)$ denote the space of functions U on Q such that

$$\tilde{U} \in \bigcap_{j=0}^{2} C^j([0, T]; H^{3-\delta-j}(\tilde{\Omega}))$$

with the obvious norm.

Now we are ready to consider the following initial-boundary value problem in Q for the Euler equations of isentropic compressible flow:

$$\begin{aligned} &\partial_t \rho + (u \cdot \nabla)\rho + \rho(\nabla \cdot u) = 0 \\ &\rho[\partial_t u + (u \cdot \nabla)u] + c^2(\rho)\nabla\rho = 0 \end{aligned} \quad \text{in } Q, \tag{2.2}$$

$$\nu_t + \nu_x \cdot 4u = 0 \quad \text{on } \Sigma, \tag{2.3}$$

$$u(x, 0) = u_0(x), \quad \rho(x, 0) = \rho_0(x) \quad \text{in } \Omega(0). \tag{2.4}$$

Here ρ is the density and $u=(u_1, u_2, u_3)$ the velocity. If $P(\rho)$ denotes the pressure, then $c^2(\rho)=dP/d\rho$ is the sound speed. We assume that $c^2(\rho)$ is a smooth, increasing function of ρ with $c^2(\rho)>0$ for $\rho>0$. The model case is that of an ideal gas in which $P(\rho)=A\rho^\gamma$, with $A>0$ and $\gamma>1$. The boundary condition (2.3) means that the velocity of the fluid is tangential to Σ. That is, fluid particles do not cross the moving boundary surface.

We shall write the system (2.2) more concisely as

$$L(U)U=0 \tag{2.5}$$

where $U=(\rho, u)$. When we wish to study the equation in the coordinates of $\tilde{Q}$, we shall write it as

$$\tilde{L}(\tilde{U})\tilde{U}=0. \tag{2.6}$$

Before stating the existence theorem, we make the following assumptions about the initial data ρ_0 and u_0:

$$\rho_0, u_0 \in H^3(\Omega(0)), \tag{2.7}$$

(2.8) There is a constant $k_0>0$ such that $\rho_0(x)\geqq k_0$ in $\Omega(0)$.

Condition (2.8) makes sense because the Sobolev inequality in three dimensions ensures that functions in $H^3(\Omega(0))$ are continuous.

We must assume a compatibility condition of the initial date with the boundary condition at $t=0$. Using the equation (2.6), and the initial data $\tilde{U}_0=(\tilde{\rho}_0, \tilde{u}_0)$, one can calculate the putative derivates,

$$\text{“}\partial_t^i\tilde{U}(0)\text{”} \quad \text{for } i=0, 1, 2,$$

as in the Cauchy–Kowaleskaya Theorem. In fact, from (2.7) and the smoothness of the nonlinear functions of $\tilde{L}$, one can deduce that “$\partial_t^i\tilde{U}(0)$”$\in H^{3-i}(\tilde{\Omega})$ for $i=0, 1, 2$. Thus it makes sense to consider the restriction of “$\partial_t^i\tilde{U}(0)$” to $\partial\tilde{\Omega}$.

For a vector field $u:\bar{Q}\to\mathbb{R}^3$ we set $(\nu_t+\nu_x\cdot u)^{\sim}=(\nu_t+\nu_x\cdot u)(\psi(y, t), t)$. Then for all $y\in\partial\tilde{\Omega}$ we assume that the data ρ_0 and u_0 satisfy

$$\partial_t^i(\nu_t+\nu_x\cdot u)^{\sim}\big|_{t=0}=0 \quad \text{for } i=0, 1, 2 \tag{2.9}$$

where we use the putative derivatives “$\partial_t^i\tilde{U}(0)$” to evaluate (2.9).

Remark. In the original coordinates of Q, the condition (2.9) can be expressed as follows:

$$\partial_\tau^i(\nu_t+\nu_x\cdot u)\big|_{t=0}=0 \quad \text{for } i=0, 1, 2$$

where $\partial_\tau=\partial_t+\partial_t\psi\cdot\nabla$ is a tangential derivative to Σ.

THEOREM 1. *Assume that the initial data satisfy* (2.7), (2.8) *and* (2.9). *Then there is a* $T>0$ *such that there is a unique classical solution* $U=(\rho, u)$ *of* (2.2), (2.3) *and* (2.4) *on* Q_T. *T depends on* k_0, *on the* H^3 *norms of* (ρ_0, u_0) *and on the derivatives of* ψ *up to order* 3.

In addition, the solution $U\in Y_{\delta,T}\cap C^1(\bar{Q}_T)$ *for each* $\delta>0$ *and*

$$\partial_t^i\tilde{U}\in L^\infty(0, T; H^{3-j}(\tilde{\Omega}))$$

for $j=0, 1, 2, 3$.

The proof of Theorem 1 is only a slight modification of that of Schochet [5] for the initial-boundary value problem in a fixed domain. For completeness, we will sketch the several steps of the proof, and indicate where changes must be made to take the noncylindrical domain into account. For details we refer the reader to [5].

The first step is to approximate the initial data (ρ_0, u_0) by functions $(\rho_0^n, u_0^n) \in H^5(\Omega(0))$ such that

$$\text{(2.10)} \qquad \begin{aligned} &\text{(i)} \quad (\rho_0^n, u_0^n) \to (\rho_0, u_0) \quad \text{in } H^3(\Omega(0)), \\ &\text{(ii)} \quad u_0^n \text{ satisfies the compatibility condition (2.9) for } i = 0, 1, 2, 3. \end{aligned}$$

This can be done by the methods of [4].

The boundary condition (2.3) is characteristic for the equations (2.2). To apply the basic existence theorem in a bounded domain for quasilinear symmetric hyperbolic systems, we must add a term to the equations (2.2) which makes the boundary condition noncharacteristic. At the same time we must add a term to the right-hand side so that the initial data ρ_0^n and u_0^n still satisfy the compatibility conditions (2.9).

According to standard results on Sobolev spaces, there exists a function $Z^n \in H^5(Q)$ such that

$$\partial_t^k(\tilde{Z}_n)(0) = \text{“}\partial_t^k \tilde{u}^n(0)\text{”} \quad \text{for } 0 \leqq k \leqq 4.$$

We then consider the modified equations (for each $\varepsilon > 0$)

$$\text{(2.2)}_\varepsilon \qquad \begin{aligned} &\partial_t \rho + (u \cdot \nabla)\rho + \rho \cdot \nabla u = 0 \\ &\rho[\partial_t u + (u \cdot \nabla)u] + c^2(\rho)\nabla\rho + \varepsilon(\nu_x \cdot \nabla)u = \varepsilon(\nu_x \cdot \nabla)Z_n \end{aligned} \qquad \text{in } Q.$$

Here we have smoothly extended the vector field ν_x into Q.

We write the system $(2.2)_\varepsilon$ more compactly as

$$\text{(2.5)}_\varepsilon \qquad L_\varepsilon(U)U = \varepsilon\gamma_n, \qquad \gamma_n = (0, (\nu_x \cdot \nabla)Z_n).$$

Next we subtract off the nonhomogeneous boundary values. Set $\beta = (0, \partial_t\psi(\phi(x, t), t))$, and let $V = (\rho, v) = U - \beta$ so that $u = v + \partial_t\psi$. Then equation $(2.5)_\varepsilon$ becomes

$$\text{(2.11)} \qquad L_\varepsilon(V + \beta)V = F(\varepsilon, n)$$

where $F(\varepsilon, n) = \varepsilon\gamma_n - L_\varepsilon(V + \beta)\beta$.

The boundary and initial conditions are

$$\text{(2.12)} \qquad \nu_x \cdot v = 0 \quad \text{on } \Sigma,$$

$$\text{(2.13)} \qquad \rho(x, 0) = \rho_0^n(x), \quad v(x, 0) = v_0^n(x) - \partial_t\psi(\phi(x, 0), 0) \quad \text{in } \Omega(0).$$

We require v to satisfy the homogeneous boundary conditions (2.12) because $v + \partial_t\psi$ should satisfy (2.4) and $\nu_t + \nu_x \cdot \partial_t\psi = 0$ on Σ. The appropriate compatibility conditions on v are

$$\text{(2.14)} \qquad \partial_t^i(\nu_x \cdot v)^{\sim}|_{t=0} = 0, \qquad i = 0, 1, 2, 3.$$

These conditions are satisfied because the initial data u_0^n satisfies (2.9) for $i = 0, 1, 2, 3$.

If we multiply the first equation of $(2.2)_\varepsilon$ by $c^2(\rho)$ and the second by ρ, the system becomes symmetric hyperbolic. Then according to the existence theorem for symmetric hyperbolic quasilinear systems with noncharacteristic boundary, for each ε and n there exists a unique solution $V(\varepsilon, n) \in X_{4, T(\varepsilon, n)}$ of (2.11)–(2.13).

Our goal is to pass to the limit as $\varepsilon \to 0$ and $n \to \infty$. However, note that $|||V(\varepsilon, n)|||_{4, T(\varepsilon, n)}$ depends on $\|F(\varepsilon, n)\|_{H^4(\tilde{Q}_{T(\varepsilon, n)})}$, where $\tilde{Q}_T = \tilde{Q} \cap \{0 < t < T\}$. This norm, in turn, involves $\|Z_n\|_{H^5(\tilde{Q}_{T(\varepsilon, n)})}$. This latter norm may blow up as $u_0^n \to u_0$ in $H^3(\Omega(0))$. Thus for each $n = 1, 2, 3, \cdots$ we choose $\tilde{\varepsilon}(n)$ so that $\|F(\tilde{\varepsilon}(n), n)\|_{H^4(\tilde{Q}_{T(\varepsilon, n)})}$ remains bounded as $n \to \infty$. Finally, we set

$$\varepsilon(n) = \min\left\{\frac{1}{n}, \tilde{\varepsilon}(n)\right\}.$$

We shall write $V(n, t)$ for $V(\varepsilon(n), n, t)$, and $T(n)$ for $T(\varepsilon(n), n)$. When it is clear from context, we may suppress the index n for brevity.

By a continuation principle of [5], $V(n) \in X_{4,T}$ for any $T > 0$ such that $|||V(n, t)|||_3$ is bounded on $[0, T)$. Hence the desired convergence will follow provided that we can show there exists a $T > 0$ with $T(n) \geqq T$ for all n, and that

$$|||V(n)|||_{3,T} \leqq C \tag{2.15}$$

where C is independent of n.

We introduce the following auxilliary norms. For $\delta > 0$ sufficiently small, we set

$$Q_\delta(0, T) = \{(x, t) \in \bar{Q}(0, T): d(y = \phi(x, t), \tilde{\Sigma}) \leqq \delta\}.$$

Next, using suitable cutoff functions equal to 1 on neighborhoods of the boundary which cover Q_δ, we use local coordinates to define $|||V|||_{m,\tan}$. This norm contains no normal derivatives at the boundary Σ. On Q/Q_δ, $|||V|||_{m,\tan}$ and $|||V|||_m$ are equivalent.

Now for $U \in X_{3,T}$, $U = (\rho, u)$, we define

$$|||U(t)|||_{E_1}^2 = |||U(t)|||_2^2 + |||U(t)|||_{3,\tan}^2 + |||\nabla \times u|||_2^2.$$

Next we cover the boundary Σ_T by a finite family of sets $\hat{G}$ such that the image of $\hat{G}$ under the mapping $(x, t) \to (\phi(x, t), t)$ is a product set $\tilde{G} \times [0, T]$ where $\tilde{G}$ is open in $\mathbb{R}^3$. Furthermore, we assume that in each $\tilde{G}$, ψ can be expressed in local coordinates

$$x_1 = y_1, \quad x_2 = y_2, \quad x_3 = y_3 + l(y_1, y_2, t).$$

Set $\alpha(x_1, x_2, x_3, t) = x_3 - l(x_1, x_2, t)$.

In these local coordinates, $\Sigma \cap \hat{G} = \{y_3 = 0\}$ and $Q \cap \hat{G} = \{y_3 > 0\}$.

Next, for $U \in X_{3,T}$, $U = (\rho, u)$ and some $\hat{G}$, we consider

$$|||\partial_{y_3}(\rho, \nabla\alpha \cdot u)|||_{2,\hat{G}}$$

where $\nabla\alpha = (-\partial_{x_1} l, -\partial_{x_2} l, 1)$ and the seminorm is evaluated in the local coordinates of $\hat{G}$. Now we sum over the sets $\hat{G}$ and define

$$|||U(t)|||_{E_2}^2 = \sum_{\hat{G}} |||\partial_{y_3}(\rho, \nabla\alpha \cdot u)|||_{2,\hat{G}}^2.$$

The basic lemma for these norms is as follows.

LEMMA 1. *Assume δ so small that the sets $\hat{G}$ cover $Q_\delta(0, T)$. Then for $U \in X_{3,T}$*

$$|||U(t)|||_3 \leqq c(|||U(t)|||_{E_1} + |||U(t)|||_{E_2}). \tag{2.16}$$

The constant is independent of U, but depends on the choice of localizing sets. The proof is given in [5].

The proof of the estimate (2.15) can be organized into several lemmas. We shall use $H_1(s), H_2(s), \cdots, K_1(s), K_2(s), \cdots$, to denote positive, smooth increasing functions of $s \geqq 0$.

LEMMA 2. *Let $V(n) \in X_{4,T(n)}$ solve* (2.11), *$\varepsilon = \varepsilon(n)$. Then for each $n = 1, 2, 3, \cdots$,*

$$\frac{d}{dt}|||V(n, t)|||_{E_1}^2 \leqq H_1(|||V(n, t)|||_3), \qquad 0 \leqq t < T(n). \tag{2.17}$$

Proof. We can show that

$$\frac{d}{dt}|||V(n, t)|||_2^2 \leqq H_2(|||V(n, t)|||_3) \tag{2.18}$$

and

$$\frac{d}{dt}|||V(n, t)|||_{3,\tan}^2 \leqq H_3(|||V(n, t)|||_3) \tag{2.19}$$

using standard estimates for symmetric hyperbolic systems. In the case of (2.18), we do not make the usual integration by parts in the spatial variables.

The remaining term in the definition of $|||\ |||_{E_1}$ is $|||\nabla \times v|||_2$. Take the curl of the second equation of (2.11) and use the fact that $\nabla \times (\nabla \rho) = 0$. We deduce that $\nabla \times v$ satisfies

$$\rho[\partial_t(\nabla \times v) + (u \cdot \nabla)(\nabla \times v)] + \varepsilon(\nu_x \cdot \nabla)(\nabla \times v) = f \tag{2.20}$$

where f is a function of first derivatives of v and ρ. Because of the boundary condition satisfied by u, the left side of (2.19) is simply a transverse (to Σ) derivative in an exterior direction. Hence usual energy type estimates show that

$$\frac{d}{dt}|||\nabla \times v|||_2^2 \leqq H_4(|||V(n, t)|||_3) \tag{2.21}$$

for $0 \leqq t < T(n)$. We combine (2.18), (2.19) and (1.21) to arrive at (2.17).

When the boundary condition is noncharacteristic, it is possible, in an equation like (2.11), to solve for derivatives of V normal to Σ in terms of tangential derivates. However, in this case the boundary matrix of the system (2.2) has rank 2, so that we can only solve for normal derivatives of two components of V in (2.11) without obtaining terms of order $1/\varepsilon$. To see this, we change variables in (2.11) using the local coordinates of $\hat{G}$ introduced earlier. The first equation becomes

$$(\nabla \alpha \cdot u + \alpha_t)\partial_{y_3}\rho + \rho\partial_{y_3}(\nabla \alpha \cdot v) = f. \tag{2.22}$$

After changing variables, and taking scalar product with $\nabla \alpha$, the second equation becomes

$$\rho(\nabla \alpha \cdot u + \alpha_t)\partial_{y_3}(\nabla \alpha \cdot v) + c^2|\nabla \alpha|^2\partial_{y_3}\rho + \varepsilon(\nu_x \cdot \nabla \alpha)\partial_{y_3}(\nabla \alpha \cdot v) = g. \tag{2.23}$$

Here f and g contain terms which involve derivatives of ρ and v in the y_1, y_2, and t directions. Now

$$\nu = (\nabla \alpha, \alpha_t)(|\nabla \alpha|^2 + \alpha_t^2)^{-1/2}$$

so that

$$\nu_x \cdot \nabla \alpha = |\nabla \alpha|^2(|\nabla \alpha|^2 + \alpha_t^2)^{-1/2}.$$

We set $w = (\rho, (\nabla \alpha \cdot v))$. Then (2.22) and (2.23) can be written

$$\Lambda \partial_{y_3} w = (f, g)$$

where

$$\Lambda = \begin{bmatrix} \nabla \alpha \cdot u + \alpha_t & \rho \\ c^2|\nabla \alpha|^2 & \rho(\nabla \alpha \cdot u + \alpha_t) + \varepsilon|\nabla \alpha|^2(\alpha_t^2 + |\nabla \alpha|^2)^{-1/2} \end{bmatrix}.$$

We can solve for $\partial_{y_3} w$ provided $\det \Lambda \neq 0$. Now on the boundary Σ, we have $\nabla \alpha \cdot u + \alpha_t = 0$ because $(u, 1)$ is tangent to Σ. Furthermore, when $t = 0$, $\rho = \rho_0^n(x) \geqq k > 0$ on $\bar{\Omega}(0)$ for n sufficiently large because convergence in $H^3(\Omega(0))$ implies uniform convergence. Hence

$$\det \Lambda = -\rho c^2(\rho)|\nabla \alpha|^2 \leqq -kc^2(k) \equiv -\lambda_0 < 0$$

on $\partial\Omega(0)$ because $|\nabla \alpha|^2 = 1 + (\partial_{x_1} l)^2 + (\partial_{x_2} l)^2 \geqq 1$. For $(y, t) \in \tilde{Q}$,

$$|\tilde{\rho}(y, t) - \tilde{\rho}_0^n(y)| \leqq t \sup_{0 \leqq s \leqq t} |\partial_t \tilde{\rho}(y, s)| \tag{2.24}$$

so that

$$\tilde{\rho}(y, t) \geqq k - ct|||V|||_{3,t}$$

where c is independent of n, n sufficiently large.

Now the data u_0^n are bounded in $H^3(\Omega(0))$ so that $\sup_{\Omega(0)} |\nabla u_0^n|$ is bounded. For each n, u_0^n satisfies the boundary condition. Thus by choosing δ_* sufficiently small we can guarantee that $\sup_{Q_{\delta_*} \cap \{t=0\}} |\nabla \alpha \cdot u_0^n + \alpha_t| \leqq \varepsilon_1$ for any suitable ε_1. Choose ε_0 and ε_1 so that for $\varepsilon \leqq \varepsilon_0$, n sufficiently large, and $x \in Q_{\delta_*} \cap \{t=0\}$,

$$\det \Lambda < -\tfrac{1}{2}\lambda_0. \tag{2.25}$$

We assume further that δ_* is chosen so small that Q_{δ_*} is covered by the local coordinate neighborhoods $\hat{G}$. Then for points in $\hat{G}$,

$$\begin{aligned} |\nabla \alpha \cdot u + \alpha_t| &\leqq \varepsilon_1 + ct \sup_{0 \leqq s \leqq t} (|\partial_t^{\sim} u| + |\partial_t \alpha_t|) \\ &\leqq \varepsilon_1 + t[C_1|||V|||_{3,t} + C_2]. \end{aligned} \tag{2.26}$$

Combining (2.24), (2.25) and (2.26) we see that

$$\det \Lambda \leqq -\tfrac{1}{2}\lambda_0 + tK(|||V(n)|||_{3,t}). \tag{2.27}$$

We have thus proved the following.

LEMMA 3. *For each n sufficiently large, there is an $S(n)$, $0 < S(n) \leqq T(n)$ such that*

$$\det \Lambda \leqq -\tfrac{1}{4}\lambda_0 \quad \textit{for } 0 \leqq t \leqq S(n) \tag{2.28}$$

on Q_{δ_*}. $S(n)$ depends on $|||V(n)|||_{3,T(n)}$.

LEMMA 4. *For each n,*

$$\frac{d}{dt}|||V(n, t)|||_{E_1}^2 \leqq K_3(|||V(n, t)|||_{E_1}) \tag{2.29}$$

for $0 \leqq t < S(n)$. K_3 does not depend on n.

Proof. On $Q_{\delta_*} \cap \{0 \leqq t < S(n)\}$ we can solve (2.22) and (2.23) for $\partial_{y_3} w$ and thus arrive at the estimate

$$|||V(n, t)|||_{E_2} \leqq K_1(|||V(n, t)|||)_E \tag{2.30}$$

for $0 \leqq t < S(n)$. For details see Schochet [5]. Now combine (2.16), (2.17) and (2.30) to deduce (2.29).

Finally, to establish the uniform bound (2.15), we will use the following argument. Assume that $S(n) = \sup\{S: \det \Lambda(n, t) \leqq -\frac{1}{4}\lambda_0 \text{ on } [0, S)\}$. $S(n) \leqq T(n)$ where $[0, T(n))$ is the maximal interval of existence of $V(n)$. Suppose that $S(n) \to 0$ as $n \to \infty$. Now the differential inequality (2.29) implies that there is a uniform bound

$$|||V(n, t)|||_{E_1} \leqq C \tag{2.31}$$

on $0 \leqq t < S(n)$. From (2.16) and (2.30), we deduce that

$$|||V(n, t)|||_3 \leqq C \quad \text{for } 0 \leqq t < S(n). \tag{2.32}$$

Finally we insert (2.32) into (2.27) to deduce that

$$\det \Lambda(n, t) \leqq -\tfrac{1}{2}\lambda_0 + tK(C_1), \qquad 0 \leqq t < S(n).$$

But then $S(n) \to 0$ as $n \to \infty$ implies that $\det \Lambda(n, t) \leqq -(3/8)\lambda_0$ on $[0, S(n))$ for n sufficiently large. By continuity, we have $\det \Lambda(t, n) \leqq -\frac{1}{4}\lambda_0$ on some larger interval, which contradicts the maximality of $S(n)$.

The proof of Theorem 1 is completed exactly as in [5], in the coordinates of $\tilde{Q}$. We use (2.11) for indices n and m, to deduce

$$\begin{aligned}\tilde{L}_n(\tilde{V}(n)+\beta)[\tilde{V}(n)-\tilde{V}(m)]&\\ =[\tilde{L}_m(\tilde{V}(m)+\beta)-\tilde{L}_n(\tilde{V}(n)+\beta)]\tilde{V}(m)+\tilde{F}(n)-\tilde{F}(m)&\\ =[\tilde{L}_m(\tilde{V}(m)+\beta)-\tilde{L}_n(\tilde{V}(n)+\beta)](\tilde{V}(m)+\beta)+\varepsilon(n)\gamma_n-\varepsilon(m)\gamma_m&\end{aligned} \tag{2.33}$$

where we have used $F(n)$ to denote $F(\varepsilon_n, n)$ when $\varepsilon=\varepsilon(n)$.

Multiplication of (2.33) by $\tilde{V}(n)-\tilde{V}(m)$ and integration over $\tilde{\Omega}$, together with the Sobolev inequality and the bound (2.15) show that $\tilde{V}(n)$ converges in $C([0, T]; L^2(\tilde{\Omega}))$. Again using the bound (2.15) and an interpolation argument, we deduce that $\tilde{V}(n)$ converges in $Y_{\delta,T}$ for each $\delta>0$. This implies that $V(n)$ converges in $C^1(\bar{Q}_T)$, and hence converges to a solution of (2.2)-(2.4). The uniqueness of C^1 solutions is well known.

3. Small amplitude motion. Now consider the case of vanishingly small boundary motion. We assume that ψ has the form

$$\psi(y, t)=y+\eta r(y, t), \tag{3.1}$$

$$\sup_{\tilde{Q}}|\partial_t r(y, t)|<\infty \tag{3.2}$$

and $\eta>0$ is a small parameter. The boundary condition can be expressed as follows:

$$\nu_x\cdot(u-\eta\partial_t r)=0 \quad \text{on } \Sigma \tag{3.3}$$

because $\nu_t=-\nu_x\cdot\partial_t\psi=-\eta\nu_x\cdot\partial_t r$.

The initial conditions are more easily described in $\tilde{\Omega}$. We take initial data of the form

$$\tilde{\rho}_0(y)=\rho_0+\eta\tilde{f}(y) \tag{3.4}$$

where ρ_0 is a constant, $\rho_0>0$, and $\tilde{f}\in H^3(\tilde{\Omega})$. For the initial velocity we assume that

$$\tilde{u}_0(y)=\eta\tilde{g}(y, \eta) \tag{3.5}$$

where $\eta\to\tilde{g}(y, \eta)$ is continuous with values in $H^3(\tilde{\Omega})$ for $0\leqq\eta\leqq\eta_0$, some $\eta_0>0$. We assume that the compatibility condition (2.9) holds for each η, $0\leqq\eta\leqq\eta_0$.

By Theorem 1, for each $\eta>0$, there will exist a solution U_η of (2.2), (2.3) with initial data (3.4), (3.5). From the estimates of Theorem 1, we can see that there will be a common interval of existence $[0, T)$, $T>0$, for all $\eta\geqq 0$. For the purpose of this section, we write the system compactly as

$$L(U_\eta)U_\eta=0 \qquad \text{in } Q_T. \tag{3.6}$$

Now let $U_0(x, t)\equiv(\rho_0, 0)$ be the constant solution with $\eta=0$, and let $L_0\equiv L(U_0)$ be the operator of (3.6) with constant coefficients. The initial-boundary value problem for the linearized acoustic equations is ($U_1=(\rho_1, u_1)$):

$$L_0U_1=0 \quad \text{in } \tilde{Q}_T, \tag{3.7}$$

$$n\cdot(u_1(y, t)-\partial_t r(y, t))=0 \quad \text{for } y\in\partial\tilde{\Omega}, \tag{3.8}$$

$$U_1(y, 0)=(\tilde{f}(y), \tilde{g}(y)) \quad \text{in } \tilde{\Omega} \tag{3.9}$$

where $\tilde{g}(y) = \tilde{g}(y, \eta = 0)$. Here we use n as the exterior unit normal to $\partial\tilde{\Omega}$. The data $\tilde{g}(y)$ satisfies the compatibility condition

$$\partial_t^i (n \cdot u_1 - \tilde{g})|_{t=0} = 0 \quad \text{for } i = 0, 1, 2. \tag{3.10}$$

The putative derivatives "$\partial_t^i u(0)$" as computed from (3.7) are used to evaluate (3.10).

We are ready to state our main result.

THEOREM 2. *Let U_η be the solution of* (3.6). *Then the solution U_1 of* (3.7), (3.8) *and* (3.9) *lies in $X_{3,T}$ and*

$$U_\eta = U_0 + \eta U_1 + o(\eta) \quad \text{as } \eta \downarrow 0. \tag{3.11}$$

The convergence takes place in $Y_{\delta,T} \cap C^1(\bar{\tilde{Q}}_T)$ for each $\delta > 0$.

Proof. To see that $U_1 \in X_{3,T}$, apply the existence theory of § 2 to the linearized equation (3.7). Since the boundary data are smooth and the compatibility condition (2.9) is satisfied, we can conclude that $U_1 \in Y_{\delta,T}$ for each $\delta > 0$ and that

$$\partial_t^j U_1 \in L^\infty(0, T, H^{3-j}(\tilde{\Omega})), \qquad j = 0, 1, 2, 3.$$

However, multiplication of the first equation by $c^2(\rho_0)$ and the second by ρ_0 yields a symmetric hyperbolic system with constant coefficients and the theory of unitary groups may be used to improve the regularity to yield $U_1 \in X_{3,T}$.

Next we define $\hat{U}_\eta$ by setting $U_\eta = U_0 + \eta \hat{U}_\eta$. Then because U_0 is constant, $\hat{U}_\eta$ satisfies:

$$L(U_\eta)\hat{U}_\eta = 0 \quad \text{in } Q_T, \tag{3.12}$$

$$\nu_x \cdot (\hat{u}_\eta - \partial_t r) = 0 \quad \text{on } \Sigma_T, \tag{3.13}$$

$$\tilde{\hat{U}}_\eta(0) = (\tilde{f}(y), \tilde{g}(y, \eta)) \quad \text{in } \tilde{\Omega}, \tag{3.14}$$

$$\partial_t^i (\nu_x \cdot (\hat{u}_\eta - \partial_t r))\tilde{}\,|_{t=0} = 0, \qquad i = 0, 1, 2 \tag{3.15}$$

where "$\partial_t^i \tilde{\hat{U}}_\eta(0)$" is computed using (3.12).

If we write (3.12) as $L(U_0 + \eta\hat{U}_\eta)\hat{U}_\eta = 0$, we can apply the quasilinear theory of § 1 to deduce that

$$\hat{U}_\eta \quad \text{is bounded in } Y_{\delta,T} \tag{3.16}$$

independent of $\eta > 0$. It follows that

$$U_\eta \to U_0 \quad \text{as } \eta \to 0 \quad \text{in } Y_{\delta,T}. \tag{3.17}$$

To prove Theorem 2, we shall show that

$$\hat{U}_\eta \to U_1 \quad \text{in } Y_{\delta,T}. \tag{3.18}$$

From (3.7) and (3.12), we have that

$$L_0(U_1 - \hat{U}_\eta) = [L(U_\eta) - L_0]\hat{U}_\eta. \tag{3.19}$$

To find the boundary conditions satisfied by $U_1 - \hat{U}_\eta$, we introduce the 3×3 matrix

$$S_\eta = (\partial_x \phi_\eta)(x, t).$$

Then $S_\eta^\tau n = \lambda \nu_x$, where $\lambda = \lambda(\eta) > 0$. The boundary condition $\nu_x \cdot (\hat{u}_\eta - \partial_t r) = 0$ on Σ can be expressed as follows:

$$n \cdot S_\eta(\hat{u}_\eta - \partial_t r) = 0 \quad \text{on } \tilde{\Sigma}_T$$

to yield

$$n \cdot (u_1 - \hat{u}_\eta) = n \cdot (S_\eta - I)\hat{u}_\eta + n \cdot (I - S_\eta)\partial_t r \quad \text{on } \hat{\tilde{\Sigma}}_T. \tag{3.20}$$

The initial condition is

$$(U_1 - \hat{U}_\eta)(0) = (0, \tilde{g}(y, \eta = 0) - \tilde{g}(y, \eta)) \quad \text{in } \tilde{\Omega}. \tag{3.21}$$

Now we need to show that the right side of (3.19) tends to zero in $L^2(\tilde{Q}_T)$ as $\eta \to 0$. We shall need to write out the matrices of L:

$$\tilde{L}(U)V = A^0(U)\partial_t V + \sum_{j=1}^{3} A^j(U)\partial_{y_j} V.$$

Recalling that $L_0 V = L(U_0)V$, we see that for each t, $0 \leqq t \leqq T$

$$\begin{aligned}\|[L_0 - \tilde{L}(\tilde{U}_\eta)]\tilde{\hat{U}}_\eta\|_{L^2(\tilde{\Omega})} &\leqq \|A^0(U_0) - A^0(\tilde{U}_\eta)\|_{L^2(\tilde{\Omega})}\|\partial_t \tilde{\hat{U}}_\eta\|_{C^0} \\ &\quad + \sum_{j=1}^{3} \|A^j(U_0) - A^j(\tilde{U}_\eta)\|_{L^2(\tilde{\Omega})}\|\partial_{y_j}\tilde{\hat{U}}_\eta\|_{C^0} \\ &\leqq \text{Const.}\, \|U_0 - \tilde{U}_\eta\|_{L^2(\tilde{\Omega})}(\|\partial_t \tilde{\hat{U}}_\eta\|_{H^{2-\delta}} + \|\partial_{y_j}\tilde{\hat{U}}_\eta\|_{H^{2-\delta}})\end{aligned}$$

for some δ, $0 < \delta < \frac{1}{2}$. Then using (3.16) and (3.17) we see that in fact $[L_0 - L(\tilde{U}_\eta)]\tilde{\hat{U}}_\eta \to 0$ in $C([0, T]; L^2(\tilde{\Omega}))$.

The same bounds and the fact that $S_\eta \to I$ uniformly on $\tilde{\Sigma}_T$ imply that the right side of (3.20) converges to zero in $C([0, T]; H^1(\partial\hat{\Omega}))$. In addition, $\tilde{g}(y, \eta) \to \tilde{g}(y, 0)$ in $H^3(\tilde{\Omega})$ by hypothesis (3.5). Then by standard results for symmetric hyperbolic systems we deduce that

$$|||\hat{U}_\eta - U_1|||_{0,T} \to 0 \quad \text{as } \eta \to 0.$$

Since $\hat{U}_\eta - U_1$ is bounded in $Y_{\delta,T}$, we deduce by the usual interpolation argument that $\hat{U}_\eta - U_1 \to 0$ in $Y_{\delta',T}$ for each $\delta' > \delta$. This completes the proof of Theorem 2.

REFERENCES

[1] J. COOPER AND W. STRAUSS, *The initial boundary problem for the Maxwell equations in the presence of a moving body*, this Journal, 16 (1985), pp. 1165–1179.

[2] A. MAJDA, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Applied Mathematical Sciences 53, Springer-Verlag, New York, 1981.

[3] A. PIERCE, *Acoustics, An Introduction to Its Physical Principles and Applications*, McGraw-Hill, New York, 1981.

[4] J. RAUCH AND F. MASSEY, *Differentiability of solutions to hyperbolic initial-boundary value problems*, Trans. Amer. Math. Soc., 189 (1974), pp. 303–318.

[5] S. SCHOCHET, *The compressible Euler equations in a bounded domain: existence of solutions and the incompressible limit*, Comm. Math. Phys., 104 (1986), pp. 49–75.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

005

AN EXISTENCE THEOREM FOR SLIGHTLY COMPRESSIBLE MATERIALS IN NONLINEAR ELASTICITY*

P. CHARRIER†, B. DACOROGNA‡, B. HANOUZET† AND P. LABORDE†

Abstract. We show that if a hyperelastic material is slightly compressible, in this case the stored energy function is a function of the "modified invariants," then the existence results of Ball are still valid. We then study the behavior of the solutions when the compressibility tends to zero.

Key words. nonlinear elasticity, slightly compressible materials, polyconvexity, coercivity

AMS(MOS) subject classifications. 49, 73

Introduction. In an important article Ball [1] has established an existence theorem for nonlinear hyperelastic and compressible materials. Roughly speaking, it is shown that the problem

$$\text{(P)} \qquad \inf\left\{ I(u) = \int_\Omega W(\nabla u(x))\, dx,\ u = u_0 \text{ on } \partial\Omega_1,\ u \in W^{1,p}(\Omega; \mathbb{R}^3) \right\}$$

admits a solution; where $\Omega \subset \mathbb{R}^3$ is the reference configuration, $\partial\Omega_1$ is part of the boundary $\partial\Omega$, $u: \Omega \to \mathbb{R}^3$ and $\nabla u \in M^3_+$ (i.e., ∇u is a 3×3 matrix with $\det \nabla u > 0$), $W: M^3_+ \to \mathbb{R}$ is the stored energy function which is assumed to be *coercive* and *polyconvex* (for precise definitions see the next section).

In particular, if the material is isotropic then it is well known that W can be written as

$$W(F) = \Phi(i(F^T F)) \tag{0.1}$$

where $\Phi: (\mathbb{R}_+)^3 \to \mathbb{R}_+$ and $i(F)$ denotes the principal invariants of F, i.e., $i(F) = (i_1(F), i_2(F), i_3(F)) = (\operatorname{tr} F, \operatorname{tr}(\operatorname{adj} F), \det F)$.

It is the aim of this article to show that the theorem of Ball still holds, under some extra hypotheses, if one replaces the principal invariants by the so-called "modified invariants" i^* (cf. [7], [8]) defined as

$$i^*(F) = (i_1^*, i_2^*) = (i_1 i_3^{-1/3}, i_2 i_3^{-2/3}). \tag{0.2}$$

Let us be more precise and first explain the importance of the "modified invariants." In practice it is a hard problem to determine experimentally the function W (or Φ, if the material is isotropic) and for slightly compressible materials, one is led to proceed indirectly. One way of determining W may be as follows: first we make some experiments (such as simple traction, biaxial traction, simple shear, etc.) for which we can assume that the volume changes are negligible,

$$W^*(F) = W(F)_{|\det F = 1} \tag{0.3}$$

or for isotropic material

$$\Phi^*(i_1, i_2) = \Phi(i_1, i_2, 1).$$

* Received by the editors November 4, 1985; accepted for publication (in revised form) December 15, 1986.

† Université de Bordeaux I, U.E.R. de Mathématiques et Informatique, 351, Cours de la Libération, 33405 Talence Cedex, France.

‡ Ecole Polytechnique Fédérale de Lausanne, Département de Mathématiques, MA (Ecublens) CH-1015 Lausanne, Switzerland.

However, since the material is actually slightly compressible, we assume that a strong pressure p leads to change of volumes and postulates that

$$p(\nabla u) = g(\det \nabla u). \tag{0.4}$$

It is then easy to show (cf. Proposition 1) that W must be in the following form:

$$W(F) = W^*((\det F)^{-1/3}F) + \int_1^{\det F} g(v)\, dv \tag{0.5}$$

or if the material is isotropic

$$\Phi(i_1, i_2, i_3) = \Phi^*(i_1^*, i_2^*) + \int_1^{\sqrt{i_3}} g(v)\, dv. \tag{0.5'}$$

However it is not obvious that such functions W (or Φ) remain *coercive* and *polyconvex* (the two crucial hypotheses in Ball's theorem). And indeed for the Mooney-Rivlin stored energy function $W^*(F) = \alpha|F|^2 + \beta|\text{adj } F|^2$, the polyconvexity is actually lost.

In the second section of this article we show (Theorem 5) that for a large class of materials including some of the Ogden materials (but not Mooney-Rivlin materials), W is coercive and polyconvex. Then, using the same techniques as in Ball [1], we show existence of minima when W satisfies (0.5).

In the last section we study the convergence of the minima when the compressibility tends to zero and, in particular, we show convergence to the incompressible case.

Finally, we should mention that this kind of rheology has been used for engineering purposes and for numerical computation of rubberlike materials; for more details we refer to Charrier and Pouyot [2] and Pouyot [9].

The article is divided as follows: § 1. Stored energy function of hyperelastic and slightly compressible materials; § 2. An existence theorem; § 3. Convergence to the incompressible case.

1. Stored energy function of hyperelastic and slightly compressible materials.

1.1. Notation. We start by recalling the usual framework of nonlinear elasticity (for references, see [10] or [3] for instance).

We denote by M^3 the set of 3×3 matrices and by

$$M^3_+ = \{A \in M^3 : \det A > 0\}. \tag{1.1}$$

We endow the space M^3 with the scalar product $A \cdot B = \text{tr}\,(AB^T)$ and we denote by $|A|$ the associated norm. We also denote by $i(A)$ the principal invariants of $A \in M^3$, i.e.,

$$i(A) = (i_1(A), i_2(A), i_3(A)) = (\text{tr } A, \text{tr (adj } A), \det A) \tag{1.2}$$

where adj A denotes the transpose of the matrix of cofactors of A.

Let $\Omega \subset \mathbb{R}^3$ be the reference configuration (Ω is a bounded open set), $u: \bar{\Omega} \to \mathbb{R}^3$ a deformation of the body satisfying $\det \nabla u > 0$, i.e., $F = \nabla u \in M^3_+$.

Let $T(u(x))$ be the Cauchy stress tensor defined on the deformed configuration $u(\bar{\Omega})$ and let $S(x)$ be the first Piola-Kirchhoff stress tensor defined as follows:

$$S(x) = \det(\nabla u(x))\, T(u(x))\, [(\nabla u(x))^T]^{-1}. \tag{1.3}$$

We also assume that the material under consideration is *hyperelastic* (and homogeneous), i.e., there exists $W: M^3_+ \to \mathbb{R}_+$ the *stored energy function* such that

$$S = W'(\nabla u) \tag{1.4}$$

where $W' \equiv (\partial W / \partial F_{i\alpha})_{1 \leqq i,\alpha \leqq 3}$.

The pressure is then equal to

$$p \equiv \tfrac{1}{3} \operatorname{tr} T = \tfrac{1}{3} (\det \nabla u)^{-1} W'(\nabla u) \cdot \nabla u. \tag{1.5}$$

If the material is also *isotropic* then W assumes the following form

$$W(\nabla u) = \Phi(i(\nabla u^T \nabla u)) \tag{1.6}$$

where $\Phi: (\mathbb{R}_+)^3 \to \mathbb{R}_+$.

As we mentioned in the Introduction, one way to construct the function Φ (or more generally W) is to determine at first

$$\Phi^*(i_1, i_2) = \Phi(i_1, i_2, 1) \tag{1.7}$$

(or $W^*(F) = W(F)_{|\det F = 1}$).

In the second step we postulate that pressure changes induce volume changes that are related in the following way:

$$p(\nabla u) \equiv g(\det \nabla u) \tag{1.8}$$

where $g: \mathbb{R}_+ \to \mathbb{R}_+$.

In Proposition 1 we show that relations (1.7) and (1.8) determine completely the function Φ (resp. W), once Φ^* (resp. W^*) and g are prescribed. And in fact we find

$$W(F) = W^*((\det F)^{-1/3} F) + G(\det F), \tag{1.9}$$

$$\Phi(i_1, i_2, i_3) = \Phi^*(i_3^{-1/3} i_1, i_3^{-2/3} i_2) + G(\sqrt{i_3}) \tag{1.9'}$$

where G is a primitive of g, i.e.,

$$G(x) = \int_1^x g(z)\, dz. \tag{1.10}$$

The identity (1.9') leads to the introduction of the so-called "*modified invariants*" (see Ogden [6], Penn [8])

$$i_1^* \equiv i_3^{-1/3} i_1 \quad \text{and} \quad i_2^* \equiv i_3^{-2/3} i_2. \tag{1.11}$$

1.2. Determination of the stored energy function. We can now state the proposition which shows (1.9) and (1.9').

PROPOSITION 1. *Part* 1. *Let* $W: M^3_+ \to \mathbb{R}_+$ *be differentiable and* $p: M^3_+ \to \mathbb{R}_+$ *be such that*

$$W'(F) \cdot F = 3 \det F p(F). \tag{1.12}$$

Let $g \in L^1_{\text{loc}}(0, +\infty)$. *The following conditions are then equivalent*:

(i) *For every* $F \in M^3_+$,

$$p(F) = g(\det F). \tag{1.13}$$

(ii) *There exists* $W^*: \{F \in M^3_+; \det F = 1\} \to \mathbb{R}_+$ *such that*

$$W(F) = W^*((\det F)^{-1/3} F) + \int_1^{\det F} g(z)\, dz. \tag{1.14}$$

Part 2. *Furthermore if* Φ *satisfies* (1.6) *then* (1.13) *is equivalent to the fact that there exists* $\Phi^*: \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *such that*

$$\Phi(i_1, i_2, i_3) = \Phi^*(i_1^*, i_2^*) + \int_1^{\sqrt{i_3}} g(z)\, dz \tag{1.14'}$$

where i_1^*, i_2^* *satisfy* (1.11).

Proof. Part 1. First suppose (ii). If W satisfies (1.14) then

$$\frac{d}{dt} W(tF)_{|t=1} = 3 \det F g(\det F)$$

and since (1.12) holds, we indeed have (1.13).

Conversely, assume that (i) holds and let

$$h(F) = W(F) - \int_1^{\det F} g(z)\, dz. \tag{1.15}$$

We then have

$$\begin{aligned} h'(F) \cdot F &= W'(F) \cdot F - (\operatorname{adj} F)^T F g(\det F) \\ &= 3 \det F(p(F) - g(\det F)) = 0. \end{aligned}$$

Therefore h is homogeneous of degree 0. Let W^* be the restriction of h on $\{F \in M^3 : \det F = 1\}$; then we immediately have (1.14). This completes Part 1.

Part 2. This is proved exactly in the same way. □

1.3. Choice of a rheology. We first start with a definition introduced by Ball [1].

DEFINITION. A function $W: M^3 \to \mathbb{R}_+$ is said to be *polyconvex* if there exists $w: M^3 \times M^3 \times \mathbb{R} \to \mathbb{R}_+$ convex, such that

$$W(F) = w(F, \operatorname{adj} F, \det F). \tag{1.16}$$

In the incompressible case a class of rheologies (i.e. stored energy functions) for isotropic materials is that of Ogden [6], where

$$W^*(F) = \sum_{i=1}^{M} a_i \operatorname{tr}(C^{\alpha_i/2}) + \sum_{i=1}^{N} b_i \operatorname{tr}((\operatorname{adj} C)^{\beta_i/2}) \tag{1.17}$$

where $C = F^T F$. If $a_i, b_i > 0$ and $\alpha_i, \beta_i \geqq 1$ then W^* defined as in (1.17) is polyconvex (cf. Ball [1]).

One can choose the coefficients in (1.17) such that W^* interpolate, with a reasonably good approximation, the experimental measurements.

We will also, marginally, consider W^* satisfying

$$W^*(F) = \sum_{i=1}^{M} a_i |F|^{\alpha_i} + \sum_{i=1}^{N} b_i |\operatorname{adj} F|^{\beta_i} \tag{1.17'}$$

with $a_i, b_i \geqq 0$ and $\alpha_i, \beta_i \geqq 1$, which is obviously polyconvex.

Note also that if $w^*: M^3 \times M^3 \to \mathbb{R}_+$ in (1.17) or (1.17') is such that

$$W^*(F) = w^*(F, \operatorname{adj} F) \tag{1.18}$$

then we have the following *coercivity* condition

$$w^*(F, G) \geqq K(|F|^\alpha + |G|^\beta) \tag{1.19}$$

where $K > 0$ is a constant and $\alpha = \max_{1 \leqq i \leqq M} \{\alpha_i\}$ and $\beta = \max_{1 \leqq j \leqq N} \{\beta_j\}$.

We now discuss the compressibility law. As we mentioned, we postulate that for slightly compressible materials the pressure p satisfies

$$p(\nabla u) = g(\det \nabla u). \tag{1.20}$$

We assume that if

$$G(x) = \int_1^x g(z)\, dz \tag{1.21}$$

then G satisfies

$$\begin{aligned} &G\colon (0,+\infty)\to[0,+\infty) \text{ is convex},\\ &\lim_{x\to 0^+} G(x)=+\infty,\\ &G(x)\geqq Cx^{\gamma} \text{ for some } C>0 \quad \text{and} \quad \gamma>1, \quad x \text{ large enough},\\ &G(x)=0 \quad \text{if and only if } x=1. \end{aligned} \tag{1.22}$$

A good example of compressibility law is, for every $F\in M^3_+$,

$$p(F)=g(\det F)=\frac{1}{2\varepsilon}\left(\det F-\frac{1}{\det F}\right), \tag{1.23}$$

which satisfies (1.22). In particular, if $\det F$ is close to 1 we have

$$p\sim\frac{1}{\varepsilon}(\det F-1)=\frac{1}{\varepsilon}\left(\frac{\rho}{\rho_0}-1\right)$$

where ρ_0 is the initial density and ρ is the density of the deformed body. This is usually considered as reasonable for numerous slightly compressible materials.

Therefore, consider a rheology of the form

$$W(F)=W^*(F)+G(\det F) \tag{1.24}$$

where W^* and G satisfy (1.17) and (1.22). We could also consider more general W^* (cf. (2.6) below). We seek deformations u such that $u=u_0$ on $\partial\Omega_1$ (where $\partial\Omega_1\subset\partial\Omega$) and which minimize the potential energy of the system (we suppose for simplicity that the material is homogeneous and that there are no external forces), i.e.,

$$I(v)=\int_\Omega W(\nabla v(x))\,dx \tag{1.25}$$

among all kinematically admissible deformations v which belong to

$$A_{\alpha\beta\gamma}=\{v\in W^{1,\alpha}(\Omega,\mathbb{R}^3),\ \operatorname{adj}\nabla v\in(L^\beta(\Omega))^9,\ \det\nabla v\in L^\gamma(\Omega)$$
$$\det\nabla v>0 \text{ a.e. and } u=u_0 \text{ on } \partial\Omega_1\}.$$

The methods and results of Ball [1] can then be applied as follows.

THEOREM (Ball). *Let $\alpha>3/2$, $\beta>1$, $1/\alpha+1/\beta<4/3$ and $\gamma>1$. If there exists $\tilde{u}\in A_{\alpha\beta\gamma}$ such that $I(\tilde{u})<\infty$ then there exists $u\in A_{\alpha\beta\gamma}$ so that*

$$I(u)=\inf\{I(v)\colon v\in A_{\alpha\beta\gamma}\}.$$

Remark 2. In the definition of $A_{\alpha\beta\gamma}$, if we assume $\alpha\geqq 2$, then there is no ambiguity in the definition of $\operatorname{adj}\nabla u$ since it is an L^1 function. However if $3/2<\alpha<2$, $\operatorname{adj}\nabla u$ is extended as a distribution by continuity denoted $\operatorname{Adj}\nabla u$ by Ball [1]. A similar remark is applicable to $\det\nabla u$.

It is the aim of this article to extend Ball's result to the case of a slightly compressible material which satisfy (cf. Proposition 1)

$$W(F)=W^*((\det F)^{-1/3}F)+G(\det F) \tag{1.26}$$

where W^* and G satisfy (1.17) and (1.22).

The theorem of Ball cannot be applied directly since, even if W^* is polyconvex and coercive when $\det F=1$, it is not, in general, the case for W satisfying (1.27). This will be investigated in the next section.

We end this section with some remarks.

Remark 3. (i) A simple example of stored energy function is the so-called Mooney-Rivlin rheology, i.e.,

$$W^*(F) = a \operatorname{tr} C + b \operatorname{tr} (\operatorname{adj} C) = a|F|^2 + b|\operatorname{adj} F|^2. \tag{1.27}$$

We shall see in the next section that, even though (1.27) is polyconvex the function

$$W(F) = a(\det F)^{-2/3}|F|^2 + b(\det F)^{-4/3}|\operatorname{adj} F|^2 + G(\det F) \tag{1.28}$$

is not polyconvex.

(ii) It is known (cf. Gurtin [4]) that in an isotropic material the constraint in the reference configuration is a pressure. The choice of a law of compressibility satisfying $G'(1) = g(1) = 0$ implies that the reference configuration is free of constraint and thus is a "natural" configuration.

2. An existence theorem.

2.1. Statement of the result. We consider a stored energy function W in the form

$$W(F) = w(F, \operatorname{adj} F, \det F) \equiv W^*((\det F)^{-1/3}F) + G(\det F) \tag{2.1}$$

where

$$W^*(F) = w^*(F, \operatorname{adj} F) \tag{2.2}$$

and w^* satisfies

$$w^*(F, H) = \sum_{i=1}^{M} a_i \operatorname{tr} (C^{\alpha_i/2}) + \sum_{j=1}^{N} b_j \operatorname{tr} (D^{\beta_j/2}) \tag{2.3}$$

with $a_i > 0$, $b_j > 0$, $C = F^TF$, $D = H^TH$ and

$$\begin{aligned} &\alpha_i \geqq \tfrac{3}{2}, \qquad i = 1, \cdots, M, \\ &\beta_j \geqq 3, \qquad j = 1, \cdots, N. \end{aligned} \tag{2.4}$$

We also assume that

$$G \text{ satisfies (1.22).} \tag{2.5}$$

Finally, if we let $\alpha = \max \{\alpha_i, 1 \leqq i \leqq M\}$ and $\beta = \max \{\beta_j, 1 \leqq j \leqq N\}$ we then assume

$$\begin{aligned} &p \equiv \frac{3\alpha\gamma}{\alpha + 3\gamma} > \frac{3}{2}, \\ &q \equiv \frac{3\beta\gamma}{2\beta + 3\gamma} \quad (\beta \geqq 3 \text{ and } \gamma > 1 \text{ imply } q > 1), \\ &\frac{1}{p} + \frac{1}{q} = \frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma} < \frac{4}{3}. \end{aligned} \tag{2.6}$$

For instance (2.6) is verified if $\alpha \geqq 2$, $\beta \geqq 3$, $\gamma > 2$.

Remark 4. The same results hold if instead of (2.3) w^* is of the form

$$w^*(F, G) = \sum_{i=1}^{M} a_i|F|^{\alpha_i} + \sum_{j=1}^{N} b_j|G|^{\beta_j}, \tag{2.3'}$$

and if we assume all the other hypotheses (2.1)-(2.6).

Recall that we want to minimize

$$I(v) \equiv \int_\Omega W(\nabla v(x))\, dx \tag{2.7}$$

over the space

$$A_{pq\gamma} = \{v \in W^{1,p}(\Omega; \mathbb{R}^3), \operatorname{adj} \nabla v \in (L^q(\Omega))^9, \det \nabla v \in L^\gamma(\Omega), \det \nabla v > 0 \text{ a.e. and } v = u_0 \text{ on } \partial\Omega_1\}. \tag{2.8}$$

The theorem can then be stated as follows.

THEOREM 5 (Existence Theorem). *Let W satisfy* (2.1)-(2.6) *and suppose that there exists $v \in A_{pq\gamma}$ such that $I(v) < \infty$, then there exists $\bar{u} \in A_{pq\gamma}$ so that*

$$I(\bar{u}) = \inf\{I(v): v \in A_{pq\gamma}\}.$$

In order to prove the above theorem, we proceed according to Ball [1] and we divide our proof in three steps. In the first one we discuss the polyconvexity of W, in the second one we study the coercivity of I in the space $A_{pq\gamma}$ and last, we pass to the limit on the minimizing sequence.

2.2. Polyconvexity of W.

PROPOSITION 6. *Let $\alpha > 0$, $\beta > 0$.*

(i) *Let*

$$W(F) = (|F|(\det F)^{-1/3})^\alpha; \tag{2.9}$$

then W is polyconvex if and only if $\alpha \geqq 3/2$.

(ii) *Let*

$$W(F) = (|\operatorname{adj} F|(\det F)^{-2/3})^\beta; \tag{2.10}$$

then W is polyconvex if and only if $\beta \geqq 3$.

(iii) *Let v_1, v_2, v_3 denote the principal stretches, i.e., the eigenvalues of $(F^TF)^{1/2}$ and let*

$$W(F) = \sum_{i=1}^{M} a_i \left(\sum_{k=1}^{3} v_k \delta^{-1/3} \right)^{\alpha_i} + \sum_{j=1}^{N} b_j \left(\sum_{j=1}^{3} v_{k+1} v_{k+2} \delta^{-2/3} \right)^{\beta_j} \tag{2.11}$$

where $\delta = \det F$ (with the notation $v_4 = v_1$ and $v_5 = v_2$) and $a_i, b_j > 0$. Furthermore if $\alpha_i \geqq 3/2$ and $\beta_j \geqq 3$ for every i, j then W is polyconvex.

Remark. It is interesting to note that, following the above proposition, the Mooney-Rivlin stored energy function

$$W^*(F) = a|F|^2 + b|\operatorname{adj} F|^2$$

does not lead to polyconvex W:

$$W(F) = a|F|^2(\det F)^{-2/3} + b|\operatorname{adj} F|^2(\det F)^{-4/3}$$

since $\alpha = 2$ and $\beta = 2$.

Proof. (i) *Step* 1. Let $h: \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ be defined as

$$h(x, \delta) \equiv (x\delta^{-1/3})^\alpha. \tag{2.12}$$

It is then easy to show that h is convex if and only if $\alpha \geqq 3/2$. Let us remark now that

$$W(F) = w(F, \det F), \quad \text{where } w(F, \delta) = h(|F|, \delta). \tag{2.13}$$

If $\alpha \geqq 3/2$, polyconvexity of W is deduced from the convexity of h and the fact that $x \to h(x, \delta)$ is a nondecreasing function.

Step 2. It now remains to show that $\alpha \geqq 3/2$ is also necessary for W to be polyconvex. To prove this is slightly more involved. We first define for $F \in M^3_+$, $a, b \in \mathbb{R}^3$ such that $F + ta \otimes b \in M^3_+$ for every $t \in \mathbb{R}_+$ (where $a \otimes b \equiv (a_i b_j)_{1 \leqq i,j \leqq 3} \in M^3$):

$$\varphi(t) \equiv W(F + ta \otimes b) = (|F + ta \otimes b| (\det (F + ta \otimes b))^{-1/3})^{\alpha}. \tag{2.14}$$

It is then easy to show (cf. Ball [1]) that

$$W \text{ polyconvex implies } \varphi \text{ convex} \tag{2.15}$$

(a function W having the property that the associated φ is convex is called *rank one convex*).

In order to simplify the notation we let $\lambda_1, \cdots, \lambda_5$ be defined as

$$\begin{aligned} |F + ta \otimes b|^2 &= \lambda_1^2 t^2 + \lambda_2 t + \lambda_3^2, \\ \det (F + ta \otimes b) &= \lambda_4 t + \lambda_5 \end{aligned} \tag{2.16}$$

for every $t \in \mathbb{R}_+$ (observe that $\det (F + ta \otimes b)$ is linear in t).

Since $W \in C^2$, so is φ and an elementary computation leads to

$$\varphi''(t) = (\lambda_1^2 t^2 + \lambda_2 t + \lambda_3^2)^{(\alpha/2)-2} (\lambda_4 t + \lambda_5)^{-(\alpha/3)-2} \left[\lambda_1^4 \lambda_4^2 t^4 \left(\frac{2\alpha}{9} (2\alpha - 3) \right) + O(t^3) \right];$$

thus

$$\varphi''(t) \geqq 0 \quad \text{for every } t \in \mathbb{R}_+ \text{ implies } \alpha \geqq \tfrac{3}{2}$$

and consequently

$$W \text{ polyconvex implies } \alpha \geqq \tfrac{3}{2}.$$

(ii) The second part of the proposition is proved in exactly the same way.

(iii) In order to prove the third part we set

$$W(F) = \psi(v_1, v_2, v_3, v_2 v_3, v_1 v_3, v_1 v_2, v_1 v_2 v_3)$$

and use a result of Ball [1] which asserts that, if $\psi: (\mathbb{R}_+)^7 \to \mathbb{R}$ is convex and increasing in each of the first six variables, then W^* is polyconvex. It is clear that if $\alpha_i \geqq 3/2$ and $\beta_j \geqq 3$ then ψ has the required properties (as in the first part). □

2.3. Coercivity of the energy functional I. We now want to show that I defined in (2.7) is coercive over the space $A_{pq\gamma}$ introduced in (2.8).

PROPOSITION 7. *Let W satisfy* (2.1)-(2.6); *then there exists $K > 0$ such that for every $v \in A_{pq\gamma}$*

$$\begin{aligned} I(v) &= \int_\Omega W(\nabla v(x))\, dx \\ &= \int_\Omega w(\nabla v, \operatorname{adj} \nabla v, \det \nabla v)\, dx \\ &\geqq K(-1 + \|\nabla v\|_{L^p}^p + \|\operatorname{adj} \nabla v\|_{L^q}^q + \|\det \nabla v\|_{L^\gamma}^\gamma). \end{aligned} \tag{2.17}$$

Remark. The result is still valid if instead of (2.3) we have (2.3′).

Proof. It is easy to see that (2.3) implies that there exists $K > 0$ such that

$$w^*(F, H) \geqq K(|F|^\alpha + |H|^\beta) \tag{2.18}$$

where $\alpha = \max \{\alpha_i : 1 \leqq i \leqq M\}$ and $\beta = \max \{\beta_j : 1 \leqq j \leqq N\}$.

Let $\delta = \det F$ and use (2.5) to get that there exists $K > 0$ (we write K generically for constants):

$$G(\delta) \geqq K(\delta^\gamma - 1); \tag{2.19}$$

hence W and w defined by (2.1) satisfy the following coercivity condition:

$$w(F, H, \delta) \geqq K(|\delta^{-1/3} F|^\alpha + |\delta^{-2/3} H|^\beta + \delta^\gamma - 1). \tag{2.20}$$

We now apply Young's inequality to

$$|F|^p = |\delta^{-1/3} F|^p \delta^{p/3}. \tag{2.21}$$

So for every $\varepsilon > 0$ and $m > 1$, $1/m + 1/m' = 1$ we have

$$|F|^p \leqq \frac{\varepsilon^m}{m} |\delta^{-1/3} F|^{pm} + \frac{\varepsilon^{-m'}}{m'} \delta^{pm'/3}. \tag{2.22}$$

Choosing p as in (2.6), i.e.,

$$p = \frac{3\alpha\gamma}{\alpha + 3\gamma}$$

and m such that $pm = \alpha$, we indeed obtain $pm' = 3\gamma$ and hence

$$|F|^p \leqq \frac{p}{\alpha} \varepsilon^{\alpha/p} |\delta^{-1/3} F|^\alpha + \frac{p}{3\gamma} \varepsilon^{-3\gamma/p} \delta^\gamma. \tag{2.23}$$

Similarly if $q = 3\beta\gamma/(2\beta + 3\gamma)$ we get

$$|H|^q \leqq \frac{q}{\beta} \varepsilon^{\beta/q} |\delta^{-2/3} H|^\beta + \frac{2q}{3\gamma} \varepsilon^{-3\gamma/2q} \delta^\gamma. \tag{2.24}$$

Combining (2.20), (2.23) and (2.24) we have indeed obtained

$$w(F, H, \delta) \geqq K(|F|^p + |H|^q + \delta^\gamma - 1) \tag{2.25}$$

and thus proved the proposition. □

Remark 8. Observe that in the above proposition we have indeed used the coercivity of $w(F, H, \delta)$ (i.e. where the arguments (F, H, δ) are *independent*) as in (2.25), and not only the condition $W(F) = w(F, \operatorname{adj} F, \det F) \geqq K(|F|^p + |\operatorname{adj} F|^q + (\det F)^\gamma - 1)$, $F \in M^3_+$ which is, in general, weaker. In fact, the coercivity of $w(F, \operatorname{adj} F, \det F)$ does not seem to be sufficient to ensure the coercivity of $I(v)$ in $A_{pq\gamma}$ if $3/2 < p < 2$, since we do not know whether $\operatorname{adj} \nabla v$ (which is defined only as a distribution in L^q) coincides with the almost everywhere definition. However, if $p \geqq 2$ and $q \geqq p/(p-1)$, the coercivity of $w(F, \operatorname{adj} F, \det F)$ is sufficient to obtain that of $I(v)$ on $A_{pq\gamma}$ (see § 2.5, Theorem 5′).

2.4. Proof of Theorem 5. We first state a lemma proved by Ball [1].

LEMMA 9. *Let Ω be a bounded open set of $\mathbb{R}^3$ and let*

$$u_k \rightharpoonup u \; W^{1,p}(\Omega; \mathbb{R}^3).$$

(i) *If $p > \frac{3}{2}$, then*

$$\operatorname{adj} \nabla u_k \rightharpoonup \operatorname{adj} \nabla u \quad \text{in } (\mathscr{D}'(\Omega))^9.$$

(ii) *Furthermore, if $p > 1$ and*

$$\operatorname{adj} \nabla u_k \rightharpoonup \operatorname{adj} \nabla u \quad \text{in } (L^q(\Omega))^9, \qquad q > 1$$

with $(1/p)+(1/q)<\frac{4}{3}$ *then*

$$\det \nabla u_k \rightharpoonup \det \nabla u \quad \text{in } \mathscr{D}'(\Omega).$$

Remark. Recall that in the lemma we have used the notation of Remark 2.

Proof of Theorem 5. We may now proceed exactly as in the proof of Ball's Theorem. Let u_k be a minimizing sequence of $I(u)$ on $A_{pq\gamma}$. A version of Poincare's inequality ensures that

$$\int_\Omega |u_k(x)|^p \leqq K\left[\int_\Omega |\nabla u_k(x)|^p + \left(\int_{\partial\Omega_1} |u_0|\, ds\right)^p\right] \tag{2.26}$$

where K is a constant. Therefore combining (2.26) and Proposition 7, we find that (up to an extraction of a subsequence)

$$\begin{aligned} &u_k \rightharpoonup \bar{u} && \text{in } W^{1,p}(\Omega;\mathbb{R}^3),\\ &\operatorname{adj} \nabla u_k \rightharpoonup H && \text{in } (L^q(\Omega))^9,\\ &\det \nabla u_k \rightharpoonup \delta && \text{in } L^\gamma(\Omega). \end{aligned} \tag{2.27}$$

The above lemma ensures that $H = \operatorname{adj} \nabla \bar{u}$ and $\delta = \det \nabla \bar{u}$. Using Proposition 6 we get

$$\liminf_{k\to\infty} I(u_k) \geqq I(\bar{u}). \tag{2.28}$$

Moreover, $\bar{u} = u_0$ on $\partial\Omega_1$ and $\det \nabla \bar{u} > 0$ a.e. Since, by (2.28), $I(\bar{u}) < +\infty$, so $\bar{u} \in A_{pq\gamma}$. Thus $\bar{u}$ is a minimizer of the energy. □

In the next section we show that the coercivity condition may be slightly improved; and in the last section of the second part of this article we will also show that one can consider more general stored energy functions.

2.5. Optimal coercivity of the energy. Recall that

$$\begin{aligned} &W(F) = w(F, \operatorname{adj} F, \det F) = W^*((\det F)^{-1/3}F) + G(\det F),\\ &W^*(F) = w^*(F, \operatorname{adj} F) = \sum_{i=1}^{M} a_i \operatorname{tr}(C^{\alpha_i/2}) + \sum_{j=1}^{N} b_j \operatorname{tr}(\operatorname{adj} C)^{\beta_j/2}, \end{aligned} \tag{2.29}$$

where $C = F^T F$,

$$G(\delta) \geqq C\delta^\gamma \quad \text{for some } C>0 \text{ and } \gamma>1,\ \delta \text{ large enough.}$$

In Proposition 7 we have shown that

$$w(F, H, \delta) \geqq K(|F|^p + |H|^q + \delta^\gamma - 1) \tag{2.30}$$

provided

$$p \leqq \frac{3\alpha\gamma}{\alpha+3\gamma}, \qquad q \leqq \frac{3\beta\gamma}{2\beta+3\gamma} \tag{2.31}$$

where $\alpha = \max\{\alpha_i : 1 \leqq i \leqq M\}$ and $\beta = \max\{\beta_j : 1 \leqq j \leqq N\}$.

In fact, we have shown that for every $(x, y, \delta) \in (\mathbb{R}_+)^3$

$$(\delta^{-1/3}x)^\alpha + (\delta^{-2/3}y)^\beta + \delta^\gamma \geqq K(x^p + y^q + \delta^\gamma - 1). \tag{2.32}$$

The question is then whether we can improve p and q by assuming that $x = |F|$, $y = |\operatorname{adj} F|$ and $\delta = \det F$. This can indeed be done and we have the following optimal result.

PROPOSITION 10. *Let W be as above, then the following two statements are equivalent*:

(i) *There exists $K>0$ such that*

$$W(F)\geqq K(|F|^p+|\text{adj } F|^q+(\det F)^r-1)$$

for every $F\in M^3_+$;

(ii) $p\leqq\max\left(\dfrac{3\alpha\gamma}{\alpha+3\gamma},\dfrac{3\beta\gamma}{\beta+6\gamma}\right)$, $q\leqq\max\left(\dfrac{3\beta\gamma}{2\beta+3\gamma},\dfrac{3\alpha\gamma}{2\alpha+6\gamma}\right)$, $r\leqq\gamma$.

Proof. Let us suppose (ii) and prove (i).

We first show the following two inequalities:

(2.33) $\text{tr}(\text{adj } A)\leqq\text{tr } A^2$ for every $A\in M^3$,

(2.34) $\text{tr } A^{1/2}\leqq\text{tr adj } A$ for every symmetric positive matrix A with $\det A=1$.

Let a_k be the eigenvalues of A. It is obvious that

$$\text{tr}(\text{adj } A)=a_2a_3+a_1a_3+a_1a_2\leqq a_1^2+a_2^2+a_3^2=\text{tr } A^2,$$

and hence (2.33). Similarly,

$$(a_2a_3)^{-1/2}+(a_1a_3)^{-1/2}+(a_1a_2)^{-1/2}\leqq a_1^{-1}+a_2^{-1}+a_3^{-1}$$

and since $\det A=1=a_1a_2a_3$ we deduce that

$$\text{tr } A^{1/2}=a_1^{1/2}+a_2^{1/2}+a_3^{1/2}\leqq a_2a_3+a_1a_3+a_1a_2=\text{tr adj } A$$

and thus (2.34).

Let us define

$$C_*=(\det F)^{-2/3}F^TF=(\det F)^{-2/3}C,$$
$$\bar\alpha=\max\left(\alpha,\frac{\beta}{2}\right),\qquad\bar\beta=\max\left(\beta,\frac{\alpha}{2}\right).\tag{2.35}$$

Then using (2.33) and (2.34) we have

$$\text{tr}(C_*^{\alpha/2})\geqq\text{tr}(\text{adj } C_*^{\alpha/4}),\qquad\text{tr}(\text{adj } C_*^{\beta/2})\geqq\text{tr}(C_*^{\beta/4})\tag{2.36}$$

and thus

$$W(F)\geqq K(\text{tr } C_*^{\bar\alpha/2}+\text{tr}(\text{adj } C_*^{\bar\beta/2})+(\det F)^\gamma-1).\tag{2.37}$$

We now proceed exactly as in Proposition 7, where we have replaced α and β by $\bar\alpha$ and $\bar\beta$, and thus we get (i). Conversely, we now assume that W satisfies the coercivity (i) and we want to show that p, q and r are as in (ii). The condition on r is immediately deduced from (i) by choosing $F=\lambda I$ and letting $\lambda\to\infty$.

Furthermore, by letting v_1, v_2, v_3 be the principal stretches of F and $\delta=v_1v_2v_3$ we get from (i)

$$\sum_{i=1}^3(v_i\delta^{-1/3})^\alpha+\sum_{i=1}^3(v_iv_{i+1}\delta^{-2/3})^\beta+\delta^\gamma$$
$$\geqq K\left[\left(\sum_{i=1}^3v_i^2\right)^{p/2}+\left(\sum_{i=1}^3v_i^2v_{i+1}^2\right)^{q/2}+\delta^r-1\right]\tag{2.38}$$

where $v_4=v_1$.

Step 1. We let s_i be such that

$$v_i = e^{s_i} \quad \text{and} \quad s_1 \geqq s_2 \geqq s_3. \tag{2.39}$$

Then (2.38) is equivalent to (considering only the leading term)

$$\begin{aligned} &e^{(s_1-(1/3)(s_1+s_2+s_3))\alpha} + e^{(s_1+s_2-(2/3)(s_1+s_2+s_3))\beta} + e^{(s_1+s_2+s_3)\gamma} \\ &\geqq K(e^{s_1 p} + e^{(s_1+s_2)q} + e^{(s_1+s_2+s_3)r} - 1). \end{aligned} \tag{2.40}$$

Hence letting

$$s_1 = s, \quad s_2 = (1-\sigma_1)s, \quad s_3 = (1-\sigma_1-\sigma_2)s$$

we must have from (2.40)

$$\begin{aligned} &\max\left\{(2\sigma_1+\sigma_2)\frac{\alpha}{3}; (\sigma_1+2\sigma_2)\frac{\beta}{3}; (3-2\sigma_1-\sigma_2)\gamma\right\} \\ &\geqq \max\{p, (2-\sigma_1)q, (3-2\sigma_1-\sigma_2)r\} \end{aligned} \tag{2.41}$$

for every $\sigma_1, \sigma_2 \geqq 0$.

Step 2. In particular

$$p \leqq \bar{p} \equiv \min_{\alpha_1,\sigma_2\geqq 0}\left\{\max_{1\leqq i\leqq 3} l_i(\sigma_1, \sigma_2)\right\} \tag{2.42}$$

where

$$l_1 = (2\sigma_1+\sigma_2)\frac{\alpha}{3}, \qquad l_2 = (\sigma_1+2\sigma_2)\frac{\beta}{3}, \qquad l_3 = (3-2\sigma_1-\sigma_2)\gamma. \tag{2.43}$$

It now remains to show that

$$\bar{p} = \max\left\{\frac{3\alpha\gamma}{\alpha+3\gamma}, \frac{3\beta\gamma}{\beta+6\gamma}\right\}$$

to obtain (ii).

We need to consider three cases.

Case 1. $\alpha \geqq 2\beta$; then $l_1 \geqq l_2$ and hence

$$\bar{p} = \min_{\sigma_1,\sigma_2\geqq 0}\{\max\{l_1, l_3\}\}.$$

Observe that the line $l_1 = l_3$ divides the cone $\sigma_1 \geqq 0$, $\sigma_2 \geqq 0$ into two regions and that $\max\{l_1, l_3\}$ tends to $+\infty$ if σ_1 or σ_2 does. Therefore, the minimum of $\max\{l_1, l_3\}$ over $\sigma_1 \geqq 0$ and $\sigma_2 \geqq 0$ is attained at one of the extremal points. Comparing the different values of $\max\{l_1, l_3\}$ at these points we find

$$\bar{p} = \frac{3\alpha\gamma}{\alpha+3\gamma}.$$

Case 2. $\alpha \leqq \beta/2$, then $l_2 \geqq l_1$ and therefore

$$\bar{p} = \min_{\sigma_1,\sigma_2\geqq 0}\{\max\{l_2, l_3\}\}.$$

Repeating the above argument we find

$$\bar{p} = \frac{3\beta\gamma}{\beta+6\gamma}.$$

Case 3. $\beta/2 \leqq \alpha \leqq 2\beta$, one shows in a similar manner that $\max\{l_1, l_2, l_3\}$ attains its minimum over $\sigma_1 \geqq 0$, $\sigma_2 \geqq 0$ exactly at the point where $l_1 = l_2 = l_3$ and this minimum is then

$$\bar{p} = \frac{3\alpha\gamma}{\alpha + 3\gamma};$$

hence the result.

Step 3. It now remains to show that

$$q \leqq \max\left(\frac{3\beta\gamma}{2\beta + 3\gamma}, \frac{3\alpha\gamma}{2\alpha + 6\gamma}\right). \tag{2.44}$$

We again use (2.41) to get that

$$2q \leqq \min_{\sigma_1 \geqq 0, \sigma_2 \geqq 0} \left\{ \max_{1 \leqq i \leqq 3} \{l_i'(\sigma_1, \sigma_2)\} \right\} \equiv 2\bar{q} \tag{2.45}$$

where $l_i' = l_i + q\sigma_1$.

A procedure similar to that of step 2 shows that $\bar{q}$ is exactly $\max(3\beta\gamma/(2\beta + 3\gamma), 3\alpha\gamma/(2\alpha + 6\gamma))$. And this concludes the proof of the proposition. □

Combining Proposition 10 with Proposition 6, we obtain the following (exactly in the same way as Theorem 5).

THEOREM 5′ (Existence Theorem). *Hypotheses* (2.1)-(2.5) *are supposed to be satisfied and* (*instead of* (2.6))

$$\begin{aligned} \bar{p} &\equiv \max\left(\frac{3\alpha\gamma}{\alpha + 3\gamma}, \frac{3\beta\gamma}{\beta + 6\gamma}\right) \geqq 2, \\ \bar{q} &\equiv \max\left(\frac{3\beta\gamma}{2\beta + 3\gamma}, \frac{3\alpha\gamma}{2\alpha + 6\gamma}\right) \geqq \frac{\bar{p}}{\bar{p} - 1} \end{aligned} \tag{2.6′}$$

holds. Assume also that there exists $v \in A_{\bar{p}\bar{q}\gamma}$ *such that* $I(v) < \infty$; *then there exists* $\bar{u} \in A_{\bar{p}\bar{q}\gamma}$ *so that*

$$I(\bar{u}) = \inf\{I(v) : v \in A_{\bar{p}\bar{q}\gamma}\}.$$

2.6. Polyconvexity of a more general stored energy function. Up to now we have only considered stored energy functions constructed from the Ogden model (i.e., W^* satisfying (2.1)-(2.3)). We now consider the more general case

$$W(F) = \psi^*(v_k^*, v_{k+1}^*, v_{k+2}^*) + G(\delta) \tag{2.46}$$

where the v_k^* are the principal stretches of $(\det F)^{-1/3}F$ and $\delta = \det F$, and $\varphi(v_k, w_k)$ stands for $\varphi(v_1, v_2, v_3, w_1, w_2, w_3)$.

For notational purposes we pose $w_k = v_{k+1}v_{k+2}$. We also make the following hypotheses on ψ^*:

$$\begin{aligned} &\psi^*: ([0, +\infty))^6 \to \mathbb{R}_+ \text{ is increasing,} \\ &\tilde{\psi}^*: (v_k, w_k) \to \psi^*(v_k^{2/3}, w_k^{1/3}) \text{ is a convex function on } [0, +\infty)^6, \\ &\psi^*(v_k, w_k) \geqq K\left(\sum_{i=1}^{3} v_i^{\alpha} + \sum_{j=1}^{3} w_j^{\beta} - 1\right) \end{aligned} \tag{2.47}$$

and G satisfies (2.5). Then if

$$p \equiv \frac{3\alpha\gamma}{\alpha + 3\gamma} > \frac{3}{2}, \quad q \equiv \frac{3\beta\gamma}{2\beta + 3\gamma} > 1, \quad \frac{1}{p} + \frac{1}{q} < \frac{4}{3}, \quad \gamma > 1 \tag{2.48}$$

we have the following.

THEOREM 5″. *If there exists* $v \in A_{pq\gamma}$ *such that* $I(v) < \infty$ *then there exists* $\bar{u} \in A_{pq\gamma}$ *such that*

$$I(\bar{u}) = \inf\{I(u): u \in A_{pq\gamma}\}.$$

Proof. The only thing that remains to be checked is the polyconvexity of W. It is then sufficient (see Ball [1]) to show that

$$\psi(v_k, w_k, \tau) = \psi^*(\tau^{-1/3} v_k, \tau^{-2/3} w_k) + G(\tau) \tag{2.49}$$

is increasing in the first six variables and is convex. The first property follows from the fact that ψ^* is increasing. It thus remains to show the convexity of ψ, but this is obvious since we have from (2.47) and (2.49)

$$\psi(v_k, w_k, \tau) = \tilde{\psi}^*((\tau^{-1/3} v_k)^{3/2}(\tau^{-2/3} w_k)^3) + G(\tau). \tag{2.50}$$

Combining Proposition 6, the convexity of $\tilde{\psi}^*$ and G, we have indeed proved the theorem. □

Examples. (i) Ogden materials, considered in the preceding sections, represent a particular case of the above theorem.

(ii) However there are examples, not of Ogden type, satisfying (2.47), for instance,

$$\psi^*(v, w) = \sum_{i,j=1}^{3} a_{ij} v_i^{\alpha_i} v_j^{\alpha_j} + b_{ij} w_i^{\beta_i} w_j^{\beta_j} + c_{ij} v_i^{\alpha_i} w_j^{\beta_j}$$

with a_{ij}, b_{ij}, $c_{ij} \geqq 0$, $\alpha_i \geqq 3/2$ and $\beta_j \geqq 3$. If we also assume that the quadratic form on $\mathbb{R}^6$

$$f(x, y) = \sum_{i,j=1}^{3} (a_{ij} x_i x_j + b_{ij} y_i y_j + c_{ij} x_i y_j)$$

is positive definite. Therefore, the convexity of f and the fact that f is increasing imply that $\tilde{\psi}^*$ is so.

3. Convergence to the incompressible case. We conclude this article by showing that if in the above analysis we let the compressibility tend to zero we recover the incompressible case.

Let $\varepsilon > 0$ and

$$I_\varepsilon(v) = \int_\Omega W^*((\det \nabla v)^{-1/3} \nabla v)\, dx + \frac{1}{\varepsilon} \int_\Omega G(\det \nabla v)\, dx \tag{3.1}$$

where W^* and G satisfy (2.1)-(2.6).

Let us consider the problem

$$\inf\{I_\varepsilon(v): v \in A_{pq\gamma}\} \tag{P_ε}$$

and define

$$I_0(v) = \int_\Omega W^*((\det \nabla v)^{-1/3} \nabla v)\, dx. \tag{3.2}$$

We define the admissible set for the incompressible case as

$$K_{\alpha\beta} = \{v \in W^{1,\alpha}(\Omega; \mathbb{R}^3), \operatorname{adj} \nabla v \in (L^\beta(\Omega))^9, \det \nabla v = 1 \text{ and } v = u_0 \text{ on } \partial\Omega_1\}. \tag{3.3}$$

We will assume that $K_{\alpha\beta} \neq \varnothing$. Finally, let

$$\inf\{I_0(v): v \in K_{\alpha\beta}\}. \tag{P_0}$$

Remark. Note that $p < \alpha$ and $q < \beta$.

We now can state our last theorem.

THEOREM 11. *Under hypotheses* (2.1)-(2.6) *and the further assumption that* $G(1) = G'(1) = 0$ *and* $G''(t) \geqq c_0 > 0$, *then every sequence* $u_\varepsilon \in A_{pq\gamma}$ *of solutions of* (P_ε) *converges* (*up to the extraction of a subsequence*) *to a solution* $\bar{u} \in K_{\alpha\beta}$ *of* (P_0) *in the following sense*:

$$
\begin{aligned}
&u_\varepsilon \rightharpoonup \bar{u} \quad \text{in } W^{1,p}(\Omega;\mathbb{R}^3),\\
&\operatorname{adj}\nabla u_\varepsilon \rightharpoonup \operatorname{adj}\nabla \bar{u} \quad \text{in } (L^q(\Omega))^9,\\
&\det \nabla u_\varepsilon \to 1 \quad \text{in } L^2(\Omega) \quad \text{strongly},\\
&\lim_{\varepsilon\to 0} I_0(u_\varepsilon) = I_0(\bar{u}),\\
&\lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\int_\Omega G(\det \nabla u_\varepsilon)\, dx = 0.
\end{aligned}
\tag{3.4}
$$

Remark. Similar results have been obtained by Pouyot [9] and Le Dret [5] for rheologies of the form $W_0(F) + (1/\varepsilon)G(\det F)$.

Proof. The existence of solutions u_ε of (P_ε) is obtained using Theorem 5; the existence of solutions u of (P_0) is a consequence of a theorem from Ball [1] for incompressible materials.

Let $w \in K_{\alpha\beta}$; then

$$
I_\varepsilon(u_\varepsilon) = I_0(u_\varepsilon) + \frac{1}{\varepsilon}\int_\Omega G(\det \nabla u_\varepsilon)\, dx \leqq I_\varepsilon(w) = I_0(w). \tag{3.5}
$$

Therefore u_ε is bounded in $A_{pq\gamma}$ and, after a possible extraction of a subsequence, we have

$$
\begin{aligned}
&\nabla u_\varepsilon \rightharpoonup \nabla \bar{u} \quad \text{in } L^p,\\
&\operatorname{adj}\nabla u_\varepsilon \rightharpoonup \operatorname{adj}\nabla \bar{u} \quad \text{in } L^q,\\
&\det \nabla u_\varepsilon \rightharpoonup \det \nabla \bar{u} \quad \text{in } L^\gamma.
\end{aligned}
\tag{3.6}
$$

Furthermore, taking into account the fact that $G(1) = G'(1) = 0$ and $G''(t) \geqq c_0 > 0$, we deduce that $c_0((t-1)^2/2) \leqq G(t)$. Hence in combination with (3.6) we get

$$
\det \nabla u_\varepsilon \to 1 \quad \text{in } L^2 \quad \text{strongly}. \tag{3.7}
$$

It remains to show that $\bar{u}$ is a solution of (P_0) and the last two equalities of (3.4). Let $\varepsilon_0 > 0$ be fixed and $0 < \varepsilon < \varepsilon_0$, we then have for $w \in K_{\alpha\beta}$

$$
I_{\varepsilon_0}(u_{\varepsilon_0}) \leqq I_{\varepsilon_0}(u_\varepsilon) \leqq I_\varepsilon(u_\varepsilon) \leqq I_\varepsilon(w) = I_0(w); \tag{3.8}
$$

thus

$$
I_0(\bar{u}) = I_{\varepsilon_0}(\bar{u}) \leqq \liminf_{\varepsilon\to 0} I_{\varepsilon_0}(u_\varepsilon) \leqq \limsup_{\varepsilon\to 0} I_{\varepsilon_0}(u_\varepsilon) \leqq I_0(w). \tag{3.9}
$$

This implies that $I_0(\bar{u}) < +\infty$ and hence $\bar{u} \in K_{\alpha\beta}$ and is a solution of (P_0). Replacing in (3.9) w by $\bar{u}$ we have immediately

$$
\lim_{\varepsilon\to 0} I_{\varepsilon_0}(u_\varepsilon) = I_0(\bar{u}) \tag{3.10}
$$

and returning to (3.8) we have indeed proved the theorem. □

REFERENCES

[1] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 3 (1977), pp. 337-407.

[2] P. CHARRIER AND J. M. POUYOT, *On the computation of rubber like materials in severe configurations*, in preparation.
[3] P. G. CIARLET, *Lecture on three dimensional elasticity*, Tata Inst. of Fundamental Research, Bombay, India, 1983.
[4] M. E. GURTIN, *An Introduction to Continuum Mechanics*, Academic Press, New York, 1981.
[5] H. LE DRET, *Incompressible behaviour of slightly compressible nonlinear elastic materials*, to appear.
[6] R. W. OGDEN, *Large deformations isotropic elasticity. On the correlation of theory and experiment for compressible rubberlike solids*, Proc. Roy. Soc. London Ser. A, 328 (1972), pp. 567–583.
[7] ———, *Volume changes associated with the deformation of rubberlike solids*, J. Mech. Phys. Solids, 26 (1978).
[8] R. W. PENN, *Volume changes accompanying the extension of rubber*, Trans. Soc. Rheology, 14 (1970).
[9] J. M. POUYOT, *Etudes numériques de problèmes d'élasticité non linéaire: application au calcul d'élastomères dans des configurations sévères*, Thèse 3ème cycle, Université de Bordeaux I, Bordeaux, France, 1984.
[10] C. TRUESDELL AND W. NOLL, *The nonlinear field theories of mechanics*, in Handbuch der Physik, Vol. III/3, S. Flügge, ed., Springer-Verlag, Berlin-New York, 1965.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

006

PERTURBATION METHODS FOR SOLID DIFFUSION IN A STEFAN PROBLEM*

JOSEPH D. FEHRIBACH†

Abstract. Consider a one-dimensional, two-phase Stefan problem where one phase is a semi-infinite solid and the second, a semi-infinite liquid, and where the dependent variable represents a diffusive impurity concentration. Assume that the diffusion coefficient for the solid phase is much less than that for the liquid and that temperature is constant in space. In this paper, singular perturbation techniques are used to study this problem when the movement of the solid-liquid interface is governed by a thermodynamic perturbation in time which is large compared to the solid diffusion coefficient. Asymptotic expansions for the solid impurity concentration are given for solids that decay, grow, or have both periods of growth and decay. It is shown that when the thermodynamic perturbations lead to decay, the boundary layer in the solid impurity concentration is substantially narrower than in the absence of the thermodynamic perturbations. The significance of this narrowing is illustrated using the liquid-phase epitaxial decay of semiconductor crystals.

Key words. Stefan problems, singular perturbation expansions, solid diffusion, liquid-phase epitaxy, crystal growth

AMS(MOS) subject classifications. primary 35R35; secondary 35K05, 41A60

1. Introduction. Perturbation methods have long been a powerful tool for studying complex systems of partial differential equations. Here these methods are applied to an infinite, one-dimensional, two-phase Stefan problem modeling "impurity" diffusion, where one phase is a semi-infinite solid and the other is a semi-infinite liquid. The problem has two perturbation sources: a thermodynamic perturbation which is time dependent and which affects the phase interface, and a singular perturbation due to the small diffusion coefficient of the solid. The interaction between these perturbations is the mathematically interesting part of the problem.

Much work has already been done on the general class of problems known as Stefan problems (e.g., [1]–[4] and their references). That the problem in question (1.1) is well posed for at least some finite time was established (up to minor adjustments) by Fasano and Primicerio [5]. But while existence and uniqueness proofs hold for all $\varepsilon > 0$, the form of the solutions to nonsingular problems with $\varepsilon \simeq 1$ differs from that of the singular problems where $\varepsilon \ll 1$. Determining and analyzing solutions for singular problems is the main goal of this paper.

In its simplest form, the problem can be stated as follows: Let $S(t)$ be the position of the solid-liquid interface with the solid phase to the left of the interface, let $v(t)$ be the interface velocity, and let u represent the concentration of the diffusing species. Then

$$u_t = \varepsilon u_{xx}, \qquad x < S(t), \tag{1.1a}$$

$$u_t = u_{xx}, \qquad x > S(t), \tag{1.1b}$$

$$\text{Stefan Condition:} \qquad v(t)[u_+ - u_-] = \varepsilon(u_x)_- - (u_x)_+, \tag{1.1c}$$

$$\text{Thermodynamic Equilibrium Condition:} \qquad F(u_-, u_+; \omega(t)) = 0, \tag{1.1d}$$

* Received by the editors January 20, 1986; accepted for publication (in revised form) October 29, 1986. This work was partially supported by the Center for Microgravity Research of the University of Alabama in Huntsville.

† Department of Mathematics, Duke University, Durham, North Carolina 27706. Current address, Department of Mathematics and Statistics, University of Alabama, Huntsville, Alabama 35899.

(1.1e) $u(x, 0) = u_s^\infty, \quad x < 0,$

(1.1f) Initial Conditions: $u(x, 0) = u_l^\infty, \quad x > 0,$

(1.1g) $S(0) = 0.$

Here ε, $\omega \ll 1$, and F is a known system of two equations that is nonsingular at $\omega(t) = 0$ and that satisfies the condition that there exist uniquely u_s^e such that $F(u_s^e, u_l^\infty; 0) = 0$. This last condition assures that the only thermodynamic influence on the motion of the interface is the perturbation $\omega(t)$. For simplicity, this thermodynamic perturbation will be assumed affine-linear in time: $\omega(t) = \delta + \alpha t$. Thus the ambient temperature is constant in space and varies slowly in time. Also the symbols + and − indicate the limiting value on the left and right sides of the interface, respectively.

This Stefan problem (1.1) can be viewed as the instantaneous heat diffusion limit of the coupled heat-impurity diffusion problem often referred to as the alloy solidification problem [2, p. 14], [6]. It should be kept in mind, however, that the presence of a single diffusion field in (1.1) places the model closer to the classical Stefan problem than the "full" alloy solidification problem. The latter problem has been studied analytically by Rubinstein [1, pp. 52–60], [7], and numerical methods for the enthalpy formulation have been given by (among others) Crowley and Ockendon [8], Fix [18], and Bermudez and Saguez [19].

Ghez and Small [9], [10] have proposed (1.1) as a model for the formation of semiconductor crystals by liquid-phase epitaxy (LPE). In an LPE process, a solid semiconductor crystal substrate is placed under a liquid semiconductor solution. Depending primarily on the ambient temperature of the solid and liquid, the crystal will either grow or decay. In the absence of a time-dependent thermodynamic perturbation (i.e., when the ambient temperature is constant), Small and Ghez [9] solved this LPE problem using the well-known similarity solution. Their solution taking the time-dependent perturbation into account, however, is only valid for very small values of time where the thermodynamic perturbation is small [10].

Notation. For convenience, let "$\tau = Or(\varepsilon)$" mean

$$\text{“}\lim_{\varepsilon \to 0} \frac{|\tau|}{|\varepsilon|} = M\text{”}$$

for some constant $M > 0$. Also let inequalities have their obvious meanings, e.g., let $\tau > Or(\varepsilon)$ imply that τ tends to zero more slowly than ε. Note that $\tau \leq Or(\varepsilon)$ is equivalent to $\tau = O(\varepsilon)$.

The solution of Ghez and Small is valid when $\omega(t) < Or(\sqrt{\varepsilon})$. In the next section, perturbation methods are used to calculate asymptotic expansions when $\omega(t) > Or(\sqrt{\varepsilon})$ (i.e., when the thermodynamic perturbation dominates) for the Stefan problem (1.1) in the growth, decay, and mixed growth-decay cases. The analysis shows that in the presence of a purely linear time-dependent perturbation (i.e., $\delta = 0$), the decaying solution contains a boundary layer, the width of which narrows from $Or(\sqrt{\varepsilon})$ to $Or(\varepsilon)$ after an initial $Or(\sqrt{\varepsilon})$ time interval. In the affine case ($\delta \neq 0$), the width of the boundary layer is $Or(\varepsilon)$ for all time. The growth solution, on the other hand, may require fractionally iterated error functions, but will always contain a transition layer of width $Or(\sqrt{\varepsilon})$ buried in the solid. The solution for the mixed growth-decay case is obtained by combining the first two solutions.

In the third section, the results of § 2 are applied to the problem posed by Ghez and Small. Here the major result is that the boundary layer in a decaying crystal becomes so narrow as to effectively have disappeared altogether.

There have been a number of other applications of asymptotic methods to Stefan problems. Many applications are discussed by Crank [2, pp. 139–162]. Among these are two: Ockendon [11] considers heat diffusion problems with either small or large latent heat, and Stewartson and Waechter [12] have studied the inward freezing of a spherical liquid assuming a large latent heat. In addition Tayler [13, pp. 167–171] discusses the difficulties inherent in an asymptotic solution to a coupled heat-impurity diffusion problem where the solid diffusivity is assumed to be zero and the liquid diffusivity is assumed small. Finally it is worth noting the classical Mullins and Sekerka [14], [15] stability analysis describing dendrite formation (see also [16] and [17]).

Weak, variational, and numerical methods have also been applied to various Stefan problems, including those modeling impurity diffusion. Among the works discussing these methods are [2, pp. 245–249], [3], [8], [9], [18], and [19].

2. The Stefan problem. When F is independent of time (i.e., $\alpha = 0$), the Stefan problem (1.1) has a similarity solution regardless of the size of ε. In particular, if $\xi = x/2\sqrt{t}$, then this solution can be written in terms of error functions:

$$u(\xi) = u_s^\infty + (u_- - u_s^\infty)\frac{\operatorname{erfc}(-\xi/\sqrt{\varepsilon})}{\operatorname{erfc}(-\lambda/\sqrt{\varepsilon})}, \qquad \xi < \lambda,$$

$$u(\xi) = u_l^\infty + (u_+ - u_l^\infty)\frac{\operatorname{erfc}(\xi)}{\operatorname{erfc}(\lambda)}, \qquad \xi > \lambda$$

where $v(t) = \lambda/\sqrt{t}$ implies that the interface position is $\xi = \lambda$. The values for u_-, u_+, and λ are determined using the Stefan condition and the requirement that $F(u_-, u_+; \delta) = 0$.

At this point one is tempted to solve the time-dependent problem (i.e., $\alpha \neq 0$) by defining $\tau = \alpha t$ (hence $\omega(\tau) = \delta + \tau$) and treating τ as a perturbation of the similarity solution. This approach is essentially the one used by Ghez and Small [10] and elsewhere [20], [21]; it works well in the liquid, but leads to difficulties in the solid for $\tau \geq Or(\sqrt{\varepsilon})$. Specifically, if τ is viewed as a perturbation of the similarity solution and $u(\xi, \tau)$ is expanded in τ, then the first order term of the solution can be written in terms of a second iterated error function. In the liquid this term is $O(\tau)$, while in the solid it is $O(\tau/\sqrt{\varepsilon})$. Therefore a different approach is required in the solid for larger times. The idea is to apply singular perturbation methods leaving $u_-(\tau)$ and $\lambda(\tau)$ as arbitrary functions to be evaluated at the interface. Note that in terms of (ξ, τ), the interface position is $\xi = \Lambda(\tau)$, where

$$\Lambda(\tau) = \frac{1}{2\sqrt{\tau}}\int_0^\tau \frac{\lambda(s)}{\sqrt{s}}\,ds, \tag{2.0}$$

which reduces to $\Lambda = \lambda$ in the similarity case.

2.1. Solid decay. First consider the case when $\omega(\tau)$ causes the solid to decay, i.e., $\lambda(\tau) < 0$. For this case, it is useful to transform the differential equation in the solid into coordinates where the position of the solid-liquid interface is fixed. Therefore define $\zeta = \xi - \Lambda(\tau)$; (1.1a) then transforms to

$$4\tau\partial_\tau u(\zeta, \tau) = \varepsilon\partial_{\zeta\zeta}u(\zeta, \tau) + 2[\lambda(\tau) + \zeta]\partial_\zeta u(\zeta, \tau). \tag{2.1}$$

Since the coefficient of $\partial_\zeta u$ is always negative, there is a boundary layer at $\zeta = 0$. The outer solution deep in the solid is simply $u_{\text{out}}(\zeta, \tau) = u_s^\infty$. To find the inner solution,

define an inner variable:

$$Z = \frac{\zeta}{\varepsilon^p}$$

where p is chosen to balance the lowest order terms in ε in (2.1). At this point two cases develop: the first will be referred to as diffusive motion, the second, thermal decay.

LEMMA 2.1. *Assume that there exists uniquely* u_s^e *such that* $F(u_s^e, u_l^\infty; 0) = 0$ *and that* $(u_l^\infty - u_s^e) = Or(\varepsilon^0)$. *If* $\omega(\tau) \leqq Or(\sqrt{\varepsilon})$, *i.e.,* δ, $\tau \leq Or(\sqrt{\varepsilon})$, *then dominant balance for* (2.1) *implies* $p = \frac{1}{2}$. *If* $\omega(\tau) = Or(\varepsilon^q)$ *for* $0 \leqq q < \frac{1}{2}$, *then* $p = 1 - q$.

Proof. Expand $\lambda(\tau) = \lambda_0 + \lambda_1 \tau + O(\tau^2)$. For diffusive motion, the Stefan condition (1.1c) and $F(u_s^e, u_l^\infty; \omega(\tau)) = O(\sqrt{\varepsilon})$ imply $\lambda_0 = Or(\varepsilon^{1-p})$ and $\lambda_1 = O(\varepsilon^0)$. On setting $p = \frac{1}{2}$, all of the terms of (2.1) scale as $Or(\varepsilon^0)$. Similar balancing of the interface conditions and (2.1) in the case of thermal decay results in $p = 1 - q$. In this latter case, the lowest order terms of (2.1) are the term containing $\lambda(\tau)$ and the second derivative term.

Remarks. (1) Little is gained by considering cases where $\delta = Or(\varepsilon^r)$ for $0 < r < \infty$ since the *coefficients* of the thermal perturbation would not be expected to depend on the solid diffusivity. Therefore from here on, either $\delta = 0$ or $\delta = Or(\varepsilon^0)$.

(2) The above lemma has the following interpretation: When the motion of the interface is governed principally by diffusion, the relative slowness of this motion leads to a boundary layer in the solid of width $Or(\sqrt{\varepsilon})$. But in the presence of a thermal perturbation which is $Or(\varepsilon^q)$ for $0 \leq q < \frac{1}{2}$, the interface motion is relatively fast, and since the solid is decaying, the boundary layer is narrowed to $Or(\varepsilon^{1-q})$. If $\delta = Or(\varepsilon^0)$, then the boundary layer has width $Or(\varepsilon)$ for all τ. When $\delta = 0$, however, there is a transition from an initial period of diffusive motion for $\tau < Or(\sqrt{\varepsilon})$ to a period of thermal decay for $\tau > Or(\sqrt{\varepsilon})$. In this case, $\lambda_0 = 0$, and the width of the boundary layer then narrows from $Or(\sqrt{\varepsilon})$ during the initial period to $Or(\varepsilon)$ for large times, i.e., $\tau = Or(\varepsilon^0)$ (cf. Fig. 2.1).

(3) The direction of the diffusive motion (growth or decay) is determined by the Stefan condition, depending on the signs of $(u_+ - u_-)$ and $(u_x)_-$. For present purposes, this direction is not important.

Now since for all τ the function $\Lambda(\tau)$ depends on the value of $\lambda(s)$ near $s = 0$ (cf. (2.0)), the initial diffusive behavior influences the large-time behavior by affecting the position of the interface in the laboratory frame (i.e., in terms of ξ or x). However, as the next lemma establishes, if distances are measured in terms of ζ, one does not need the solution for the diffusive period to calculate the solution for the thermal period. The lemma follows immediately from the change of variables.

LEMMA 2.2. *Recall that* $\zeta = \xi - \Lambda(\tau)$. *When written in terms of* (ζ, τ), *for* $\tau > O(\sqrt{\varepsilon})$, *system* (1.1) *is independent of the initial diffusive period.*

A lowest order asymptotic solution valid for $\omega(\tau) > Or(\sqrt{\varepsilon})$ can now be calculated using a multiple time scaling. Recall that $\omega(\tau) = \delta + \tau$, and assume first that $\delta = 0$ and $\tau = Or(\varepsilon^q)$ for $0 \leqq q < \frac{1}{2}$. The correct inner variables in the solid are $Z = \zeta/\varepsilon^{1-q}$ and $\tilde{\tau} = \tau/\varepsilon^q$. Let $r = 1 - 2q$. On writing (2.1) in terms of $(Z, \tilde{\tau})$, one obtains the inner equation

$$4\varepsilon^r \tilde{\tau} \partial_{\tilde{\tau}} u_{\text{in}}(Z, \tilde{\tau}) = \partial_{ZZ} u_{\text{in}}(Z, \tilde{\tau}) + 2[\lambda(\tilde{\tau}\varepsilon^q) + \varepsilon^r Z] \partial_Z u_{\text{in}}(Z, \tilde{\tau}). \tag{2.2}$$

Now expand $u_{\text{in}}(Z, \tilde{\tau}) = u_{\text{in}}^0(Z, \tilde{\tau}) + O(\varepsilon^r)$. A priori $\lambda(\tilde{\tau})$ would also be expected to depend on ε. But because the interface conditions written in terms of the inner variables in the decay case are independent of ε (cf. (1.1c), (1.1d)), $\lambda(\tilde{\tau})$ is also independent of ε. On substituting the expansion for u_{in} into (2.2), solving the lowest order differential equation for u_{in}^0 and matching this solution with the outer solution deep in the solid,

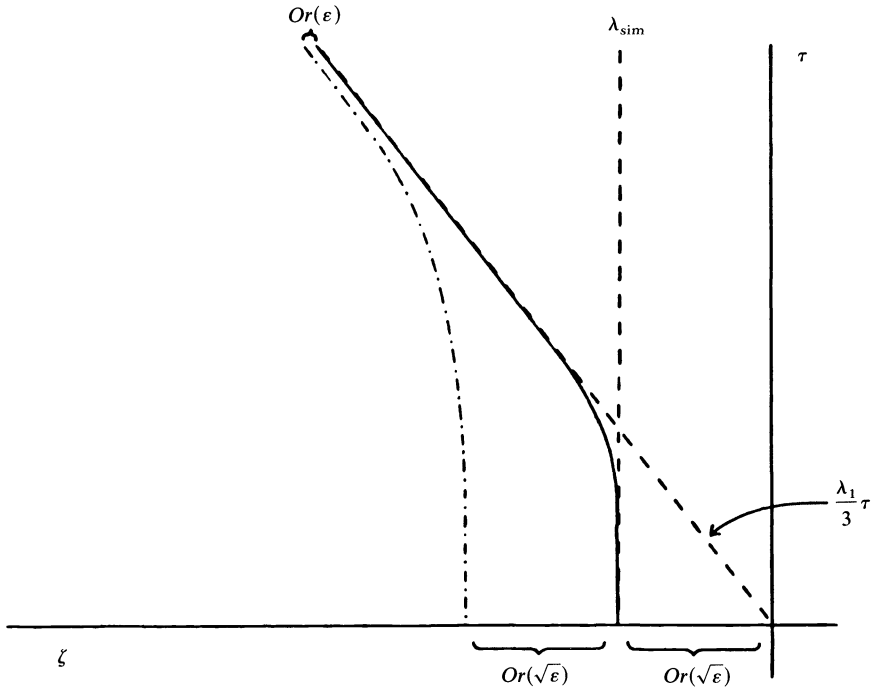


FIG. 2.1. *Transition from diffusive decay to thermal decay. The continuous line is the position of the interface and the dash-dot line shows the thickness of the boundary layer.* λ_{sim} *is the value of* λ *in the similarity solution, and* $(\lambda_1/3)\tau$ *is the interface position for large times.*

one obtains a lowest order solution $u^0(\zeta, \tau)$ valid uniformly in ζ for $\tau = Or(\varepsilon^q)$. On the other hand, if $\delta = Or(\varepsilon^0)$, the scaling of τ is not important and the above perturbation expansion can be carried out with $p = 1$. In either case, the following lemma is obtained.

LEMMA 2.3. *Assume that* $\omega(\tau) \geqq Or(\sqrt{\varepsilon})$. *Then to lowest order, the uniform (in* ζ*) solution of* (2.1) *is*

$$u^0(\zeta, \tau) = u_s^\infty + [u_-(\tau) - u_s^\infty]\, e^{-2\lambda(\tau)\zeta/\varepsilon} \tag{2.3}$$

where the functions $u_-(\tau)$ *and* $\lambda(\tau)$ *are determined by linearizing the interface conditions in* τ *(recall that F is assumed to be nonsingular).*

Note that even for $\delta = 0$, this lowest order solution is independent of q. The order of the next term in the expansion, however, is $1-2q$. Also note that the exponential factor in (2.3) would be present in every term of the expansion, regardless of the order.

2.2. Solid growth. Now consider the case when $\omega(\tau)$ causes the solid to grow, i.e., $\lambda(\tau) > 0$. In terms of (ξ, τ), (1.1a) is

$$4\tau \partial_\tau u(\xi, \tau) = \varepsilon \partial_{\xi\xi} u(\xi, \tau) + 2\xi \partial_\xi u(\xi, \tau). \tag{2.4}$$

Since the coefficient of $\partial_\xi u$ in (2.4) vanishes at $\xi = 0$, in the growth case a transition layer is buried in the growing solid. As in the decaying case, the goal is to find an asymptotic expansion for u in the solid for $\omega(\tau) > O(\sqrt{\varepsilon})$. But as before, if $\delta = 0$ there is an initial period when the motion of the interface is governed by diffusion. Since $\Lambda(\tau)$ depends on the value of λ in this initial period, and since for a growing solid $\Lambda(\tau)$ is the distance between the transition layer and the solid-liquid interface, it is

not possible to move to a reference frame where calculations can be made completely independent of the initial period. Since the length of this initial period is $O(\sqrt{\varepsilon})$, however, lowest order calculations can be made ignoring this period. Thus for large times, the position of the interface is

$$\xi = \Lambda(\tau) = \lambda_0 + \frac{\lambda_1}{3}\tau + O(\sqrt{\varepsilon}, \tau^2)$$

where λ_0 and λ_1 are determined by the interface conditions (1.1c, d) for $\tau > O(\sqrt{\varepsilon})$. Note that if $\delta = Or(\varepsilon^0)$, there is no diffusive period and therefore no $O(\sqrt{\varepsilon})$ error term.

Finding a uniform asymptotic solution for the solid in the growth case implies combining two outer solutions, one for $\xi < 0$, one for $0 < \xi < \Lambda(\tau)$, with an inner solution at $\xi = 0$. As before, the outer solution deep in the solid is simply $u_{\text{left}}(\xi, \tau) = u_s^\infty$.

Now consider the region between the transition layer and the interface, and let $u_r \equiv u_{\text{right}}$ be the outer solution in this region. To lowest order in ε, the outer equation corresponding to (2.4) is

$$2\tau \partial_\tau u_r^0(\xi, \tau) = \xi \partial_\xi u_r^0(\xi, \tau). \tag{2.5}$$

This equation simply indicates that u_r^0 is constant along the characteristic curves $\xi^2\tau = x = \text{constant}$. Let u_-^{00} and u_-^{01} be the interface constants for the solid solution, i.e., require that

$$u_r^0(\Lambda(t), t) = u_-^{00} + u_-^{01} t + O(t^2). \tag{2.6}$$

Then to lowest order in ε, $u_r(\xi, \tau)$ is found by solving $\xi^2\tau = (\Lambda(t))^2 t$ for t and substituting into (2.6). If $\Lambda(t)$ is approximated to $O(t^2)$, then the equation to be solved is a cubic with one real root. The resulting outer solution is then

$$u_r^0(\xi, \tau) = u_-^{00} + u_-^{01}[Q_+(\xi, \tau) + Q_-(\xi, \tau)]^2 + \text{HOT} \tag{2.7}$$

where HOT represents higher order terms and

$$Q_\pm(\xi, \tau) = \left[\frac{3\xi\sqrt{\tau}}{2\lambda_1} \pm \left[\frac{\lambda_0^3}{\lambda_1^3} + \frac{9\xi^2\tau}{4\lambda_1^2}\right]^{1/2}\right]^{1/3}. \tag{2.8}$$

Note that if $\delta = 0$, then $\lambda_0 = 0$ and (2.7) and (2.8) reduce to

$$u_r^0(\xi, \tau) = u_-^{00} + u_-^{01}\left[\frac{9\xi^2\tau}{\lambda_1^2}\right]^{1/3} + O(\tau^{2/3}).$$

Again the coefficients u_-^{00}, u_-^{01}, λ_0 and λ_1 are found using the interface conditions.

Now consider the inner equation near $\xi = 0$. By dominant balance, the appropriate inner equation is

$$4\tau \partial_\tau u_{\text{in}}(X, \tau) = \partial_{XX} u_{\text{in}}(X, \tau) + 2X \partial_X u_{\text{in}}(X, \tau) \tag{2.9}$$

where $X = \xi/\sqrt{\varepsilon}$. To lowest order in ε, the inner solution is essentially the similarity solution

$$u_{\text{in}}^0(X, \tau) = u_s^\infty + \tfrac{1}{2}[u_-^{00} - u_s^\infty] \operatorname{erfc}(-X).$$

This lowest order inner solution, however, may not be sufficient to construct a uniform solution which is $O(\varepsilon)$ or $O(\sqrt{\varepsilon})$ for all ξ since fractional powers of ε (other than $\sqrt{\varepsilon}$) may appear in the expansion. To determine if any such powers are present, the following lemma is needed. The lemma shows the effect on the inner solution of a nonconstant outer solution u_r satisfying (2.5). In particular note that higher order terms from the outer solution do not affect the lower order terms of the inner solution.

LEMMA 2.4. *Suppose that $u_{\text{in}}(X, \tau)$ is an inner solution satisfying* (2.9) *and suppose that u_-^{00}, u_s^{∞}, K, m, n and q are constants with $0 < q < m+n$. If $u_{\text{in}}(X, \tau)$ matches the outer solutions $u_{\text{left}}(\xi, \tau) = u_s^{\infty}$ on the left and $u_r(\xi, \tau) = u_-^{00} + K(\xi^2\tau)^q + O(\tau^m \varepsilon^n)$ on the right, then*

$$u_{\text{in}}(X, \tau) = u_s^{\infty} + \tfrac{1}{2}[u_-^{00} - u_s^{\infty}] \operatorname{erfc}(-X) + Kq(\varepsilon\tau)^q \operatorname{ierfc}^{2q}(-X) + O(\varepsilon^{m+n}, \varepsilon)$$

where $\operatorname{ierfc}^{2q}$ *is defined using an extension of the Dirichlet formula for repeated integrals*:

$$\operatorname{ierfc}^{2q}(X) \equiv \int_0^{\infty} s^{2q-1} \operatorname{erfc}(s+X)\, ds.$$

Remark. The function $\operatorname{ierfc}^{2q}$ can be thought of as a fractionally iterated error function. For $2q \in \mathbb{Z}^+$, it agrees (up to a factor of $\Gamma(2q)$) with the usual definition for an iterated error function. The function equivalently can be viewed as a convolution of $\operatorname{erfc}(X)$ with the generalized function x_+^{λ}. The definition can then be extended to include fractional derivatives when $q \leqq 0$ [22, p. 48].

Proof. Confirming that u_{in} satisfies the differential equation (2.9) is a simple application of integration by parts. Since the latter two terms in u_{in} both vanish as $X \to -\infty$, u_{in} correctly matches u_{left}. On the right, the sum of the first two terms asymptotically tends to u_-^{00}. As for the third term, fix β so that $0 \ll \beta \ll X$ and let $p = 2q - 1$. Then since $\operatorname{erfc}(-s) = 2 - \operatorname{erfc}(s)$,

$$\begin{aligned}\int_0^{\infty} s^p \operatorname{erfc}(s-X)\, ds &= \int_0^{2X} s^p \operatorname{erfc}(s-X)\, ds + \text{TS} \\ &= \frac{1}{q} X^{2q} + \int_0^{X} [(X+s)^p - (X-s)^p] \operatorname{erfc}(s)\, ds + \text{TS}\end{aligned}$$

where TS represents terms that are transcendentally small, i.e., terms which, as $X \to \infty$, tend to zero faster than X^{-n} for all n. But

$$\begin{aligned}\int_0^{X} [(X+s)^p - (X-s)^p] \operatorname{erfc}(s)\, ds &\leqq X^p \int_0^{\beta} \left[\left(1+\frac{t}{X}\right)^p - \left(1-\frac{t}{X}\right)^p\right] dt + \text{TS} \\ &\leqq pX^{p-1}\beta^2 + O(X^{p-3}).\end{aligned}$$

Hence for $X \gg 0$, $Kq(\varepsilon\tau)^q \operatorname{ierfc}^{2q}(-X)$ asymptotically approaches $K(\xi^2\tau)^q$. Finally terms which are $O(\varepsilon^n)$ in the outer solution match similar terms in the inner solution, while $O(\tau^m)$ terms in the outer solution correspond to $O(\varepsilon^m)$ terms in the inner solution. Also the term $\varepsilon \partial_{\zeta\zeta} u$ in (2.4) implies that the size of the next terms in the inner expansion will be at least $O(\varepsilon)$.

To construct a uniform solution for the growth case, the outer solutions must now be analyzed on the scale of the inner variable and matched to the appropriate inner solution. Let $u_m(X, \tau)$ be the right outer solution written on the scale of the inner variable. Two cases again arise: first when $\delta = 0$ so that $\lambda_0 = 0$, and second when $\delta = Or(\varepsilon^0)$ so that $\lambda_0 = Or(\varepsilon^0)$ also. In the first case, $u_m(X, \tau) = u_r(\xi, \tau)$, i.e.,

$$u_m(X, \tau) = u_-^{00} + u_-^{01} \varepsilon^{1/3} \left[\frac{9X^2\tau}{\lambda_1^2}\right]^{1/3} + O(\varepsilon^{2/3}), \qquad X > 0.$$

In the second case, in terms of the inner variable, (2.8) becomes

$$Q_{\pm}(\xi, \tau) = \pm\left[\frac{\lambda_0}{\lambda_1}\right]^{1/2} + \sqrt{\varepsilon}\, \frac{X\sqrt{\tau}}{2\lambda_0} + O(\varepsilon);$$

hence

$$u_m(X,\tau)=u_-^{00}+u_-^{01}\frac{\varepsilon X^2\tau}{\lambda_0^2}+O(\varepsilon^{3/2}),\qquad X>0.$$

Uniform solutions for the solid concentration can now be obtained by adding the inner and outer solutions, then subtracting the matching terms. When $\delta=0$, the inner solution is given by the Lemma 2.4 with $q=\frac{1}{3}$. The uniform solution for $\tau>Or(\sqrt{\varepsilon})$ in this case is thus

$$u(\xi,\tau)=u_s^\infty+\tfrac{1}{2}[u_-^{00}-u_s^\infty]\operatorname{erfc}(-\xi/\sqrt{\varepsilon})+u_-^{01}\left[\frac{\varepsilon\tau}{3\lambda_1^2}\right]^{1/3}\operatorname{ierfc}^{2/3}(-\xi/\sqrt{\varepsilon})+O(\sqrt{\varepsilon},\tau^{2/3}).$$

When $\delta=Or(\varepsilon^0)$, $q=1$. Hence in this case the terms of the inner solution which match the lowest order terms in τ and ε of the outer solution are higher order in ε. Therefore the uniform solution is

$$u(\xi,\tau)=u_s^\infty+\tfrac{1}{2}[u_-^{00}-u_s^\infty]\operatorname{erfc}(-\xi/\sqrt{\varepsilon})+u_-^{01}S(\xi,\tau)+O(\varepsilon,\tau^2)$$

where

$$S(\xi,\tau)=\begin{cases}0, & \xi\leqq 0,\\ [Q_+(\xi,\tau)+Q_-(\xi,\tau)]^2, & \xi>0.\end{cases}$$

2.3. Mixed growth-decay problems. Up until this point, in all of the problems analyzed, the thermal perturbation has caused either growth or decay for all values of τ. In this section attention will be given to mixed problems where the time-dependent and time-independent terms of $\omega(\tau)$ drive the interface in opposite directions.

Consider the case when the time-independent term induces growth while the time-dependent term induces decay (a similar argument can be given if the signs on the terms are reversed). Under these conditions, the crystal will initially grow since $\lambda_0>0$, but after a time the process will reverse to decay since $\lambda_1<0$. The turning point occurs when $\lambda(\tau)=0$ which (to $O(\sqrt{\varepsilon},\tau^2)$) corresponds to $\tau=\tau_{tp}\equiv-(\lambda_0/\lambda_1)$ and $\xi=\Lambda(\tau_{tp})=\frac{2}{3}\lambda_0$. The mixed solution is then found by combining a growth solution (similar to that of § 2.2) for $0\leqq\tau\leqq\tau_{tp}$ with a decay solution (similar to that of § 2.1) for $\tau\geqq\tau_{tp}$. The accuracy of these inequalities, however, is limited to $O(\sqrt{\varepsilon})$ since the rapid growth/decay approximation is not valid for an $O(\sqrt{\varepsilon})$ time interval around the turning point. Indeed this accuracy limit holds for all equalities and inequalities throughout this section.

The growth solution is again constant along the characteristic curves $\xi^2\tau=$ constant and again is found by solving $(\Lambda(t))^2t=\xi^2\tau$ for t and substituting into (2.6). In this case, however, when $\Lambda(\tau)$ is approximated to $O(t^2)$, the resulting cubic has three real roots. The appropriate root is the smallest of the three; using this value for t, one finds that the uniform solution in the solid for $0\leqq\tau\leqq\tau_{tp}$ is

$$u(\xi,\tau)=u_s^\infty+\tfrac{1}{2}[u_-^{00}-u_s^\infty]\operatorname{erfc}(-\xi/\sqrt{\varepsilon})+u_-^{01}R(\xi,\tau)+O(\varepsilon,\tau^2)\tag{2.10}$$

where u_-^{00} and u_-^{01} are determined as before by the interface conditions and where

$$R(\xi,\tau)=\begin{cases}0, & \xi\leqq 0,\\ \tfrac{1}{4}[Q_+(\xi,\tau)+Q_-(\xi,\tau)+\sqrt{-3}[Q_+(\xi,\tau)-Q_-(\xi,\tau)]]^2, & \xi>0.\end{cases}$$

Note that since in this case Q_+ and Q_- are complex conjugates, $R(\xi,\tau)$ is real and continuous at $\xi=0$. Also note that $-(\lambda_0/\lambda_1)$ is not only the value of τ for which $\lambda(\tau)=0$, but is also the maximum value of τ for which (2.4) can be solved using characteristics when λ_1 is negative. The solution for $\tau\leqq\tau_{tp}$ is illustrated in Fig. 2.2(a).

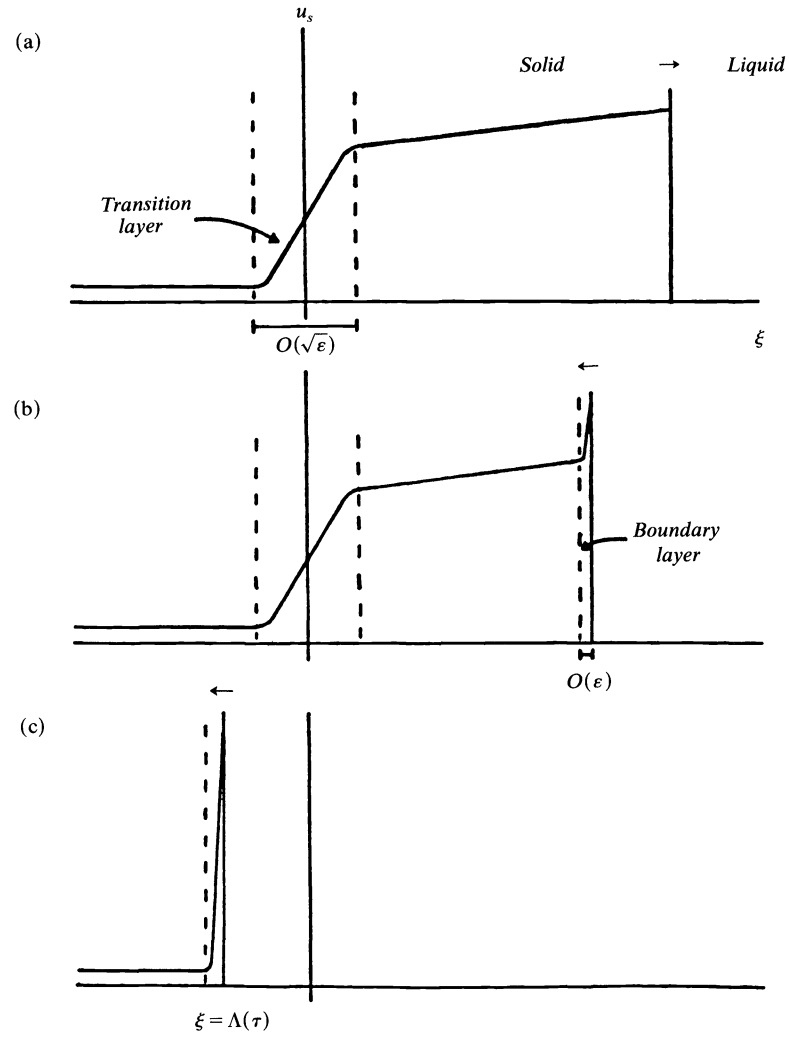


FIG. 2.2. *Mixed growth-decay case: When subject to both undercooling* ($T_0 < T_{sat}$) *and a temperature ramp* ($\sigma = 1$), (a) *the crystal initially grows,* (b) *then begins to decay,* (c) *finally resembling pure decay.*

At $\tau = \tau_{tp}$ growth ends and decay begins. Since $u(\xi, \tau_{tp})$ in the liquid is not constant, the decay solution of the previous section is not immediately applicable. The variation in u is $O(\tau_{tp})$, however, and can be viewed as a new perturbation of the interface concentrations. Therefore the formula for $u(\xi, \tau)$ in the solid is still determined by (2.4) using essentially the same argument as in § 2.1, except that the inner solution must now be matched to the growth solution given by (2.10) rather than to u_s^∞. The uniform solution for the solid valid beginning at $\tau = \tau_{tp}$ is thus

$$u(\xi, \tau) = u_s^\infty + \tfrac{1}{2}[u_-^{00} - u_s^\infty] \operatorname{erfc}(-\xi/\sqrt{\varepsilon}) + u_-^{01} R(\xi, \tau)$$
$$+ [u_{d-}^0(\tau) - u_-^{00} - u_-^{01} R(\Lambda(\tau), \tau)] 4 e^{-2\lambda_d(\tau)(\xi - \Lambda(\tau))/\varepsilon} + O(\sqrt{\varepsilon}, \tau^2)$$

where for $\tau > \tau_{tp}$, the coefficient functions $u_{d-}(\tau)$ and $\lambda_d(\tau)$ are respectively the interface concentration for the solid and the interface velocity during the decay period. These

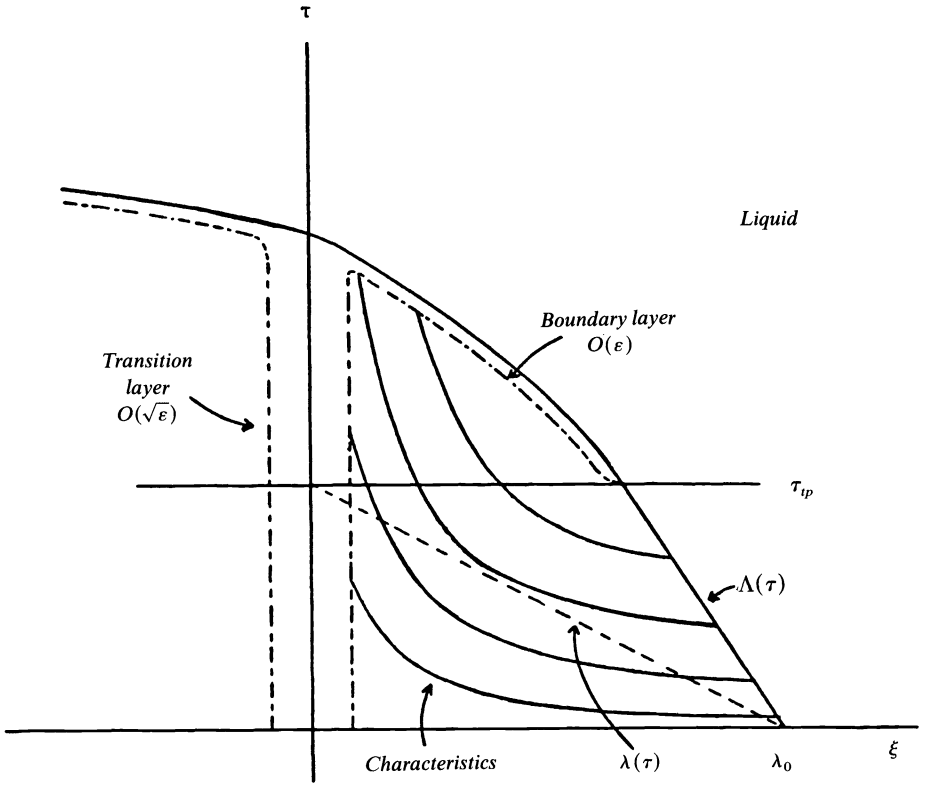


FIG. 2.3. *The interface, transition and boundary layers in the $\xi\tau$-plane.*

functions can be determined using the interface conditions and taking the new perturbation into account. Figure 2.2(b) illustrates this solution.

Let τ_0 be the value of τ such that $\Lambda(\tau_0) = 0$. For τ near τ_0 (i.e., for $\tau = \tau_0 \pm O(\sqrt{\varepsilon})$), the retreating interface moves across the transition layer. After this interval, the solution in the solid is then very similar to the solution given in the pure decay case (Fig. 2.2(c)): the general profiles of these solutions are the same, and the exact interface coefficients differ only due to the boundary perturbations in the liquid phase. Indeed because of the semi-infinite nature of this phase, the significance of these perturbations decreases with time, and this solution asymptotically approaches the solution of the pure decay case.

The regions for which the various solutions are valid are shown in Fig. 2.3. Note the relative widths of the transition and boundary layers.

3. The semiconductor problem. The results of the previous section will now be used to study the effects of a linear temperature ramp on the LPE decay of semiconductor crystals (applications to LPE growth or mixed growth-decay problems can also be made). This problem is somewhat more complicated than the one studied in § 2 since there are now several unknown concentrations to be found in each phase. Hence (1.1) is transformed to a coupled system of Stefan problems. To simplify notation and because of the particular interest in gallium-aluminum-arsenide (GaAlAs), the present discussion will focus on this semiconductor. It should be noted, however, that the results given here are applicable to any one-dimensional LPE process.

The principal assumptions in this model are that the problem is spacially one dimensional and that each phase is semi-infinite. Since (1.1) is classically well posed for sufficiently small thermal perturbations [5], there is no "mushy region" in the model. Thus the formation of dendrites or pits in the solid is not taken into account. Since GaAlAs crystals formed by LPE typically "grow smooth but decay rough" (cf. [23]-[25]), a model incorporating pit formation in the decaying case would be preferred. However, the physical implications of this model are still interesting. Also the semi-infinite assumption for the liquid is valid only for an initial period since the liquid is in fact of finite depth. (The depth of the solid is also finite, but the diffusion coefficient in this phase is so extremely small that the semi-infinite assumption for this phase is valid for any practical time.) Since numerical methods can only easily handle times outside of this initial period, however, the solution obtained by assuming a semi-infinite liquid can be used as a "starter solution" for an analytic-numerical scheme valid for a wide range of times [9].

The crystal structure for GaAlAs is face-centered cubic (or diamond) with tetrahedral bonding. Half of the lattice sites are occupied by arsenic atoms, the other half by either gallium or aluminum atoms. For the present model the aluminum atoms are viewed as diffusing across the gallium sites. In the semiconductor liquid, gallium serves as the solvent with aluminum and arsenic present in substantially smaller amounts. Hence aluminum and arsenic are viewed as the diffusing species in this phase.

Let the concentrations of the constituent species be measured in mole fractions for both the solid and liquid phases. Specifically, let u_i be the fraction of liquid made of the ith constituent and $u_{i,s}$ the fraction of the solid made of the ith constituent. To simplify notation, let the aluminum fraction of the solid be denoted u_s. Finally let D_s, D_{Al} and D_{As} be the diffusion coefficients for aluminum in the solid and aluminum and arsenic in the liquid. Equations (3.1) then govern these concentrations in both phases.

Partial Differential Equations:

Solid: $x < S(t)$:

$$u_{\mathrm{Ga},s} + u_s = \tfrac{1}{2}, \tag{3.1a}$$

$$\partial_t u_s(x, t) = D_s \partial_{xx} u_s(x, t), \tag{3.1b}$$

$$u_{\mathrm{As},s} = \tfrac{1}{2}. \tag{3.1c}$$

Liquid: $x > S(t)$:

$$u_{\mathrm{Ga}} + u_{\mathrm{Al}} + u_{\mathrm{As}} = 1, \tag{3.1d}$$

$$\partial_t u_{\mathrm{Al}}(x, t) = D_{\mathrm{Al}} \partial_{xx} u_{\mathrm{Al}}(x, t), \tag{3.1e}$$

$$\partial_t u_{\mathrm{As}}(x, t) = D_{\mathrm{As}} \partial_{xx} u_{\mathrm{As}}(x, t). \tag{3.1f}$$

Equations (3.1) are coupled at the solid-liquid interface by two Stefan conditions, one for arsenic, one for aluminum, and two thermodynamic equilibrium conditions that represent the continuity of chemical potentials across the phase interface and that are determined by the phase diagram for GaAlAs. As before, the initial conditions are that all of the concentrations be constant and that $S(0) = 0$.

Interface Conditions: $x = S(t)$:

Stefan Conditions:

$$v(t)[u_s(t) - u_{\mathrm{Al}}(t)] = D_{\mathrm{Al}} \partial_x u_{\mathrm{Al}}(t) - D_s \partial_x u_s(t), \tag{3.2a}$$

$$v(t)[\tfrac{1}{2} - u_{\mathrm{As}}(t)] = D_{\mathrm{As}} \partial_x u_{\mathrm{As}}(t). \tag{3.2b}$$

Equilibrium Conditions:

$$\mu_{Ga}(T, u) + \mu_{As}(T, u) = \mu_{GaAs}(T, u), \tag{3.2c}$$

$$\mu_{Al}(T, u) + \mu_{As}(T, u) = \mu_{AlAs}(T, u). \tag{3.2d}$$

Initial Conditions:

$$u_s(x, 0) = u_s^{\infty},$$
$$u_{Al}(x, 0) = u_{Al}^{\infty},$$
$$u_{As}(x, 0) = u_{As}^{\infty}.$$

The initial aluminum concentration in the solid is often (though not always) taken to be zero. The initial liquid concentrations are determined by the process used to prepare the liquid. Assume that the liquid is prepared at a saturation temperature T_{sat} to be uniformly in equilibrium with a solid whose aluminum concentration is u_s^e.

In (3.2c), (3.2d), μ_i represents the chemical potential of the ith species as a function of temperature and constituent concentrations. Assume that the temperature is ramped, i.e., let $T(t) = T_0 + rt$, where T_0 is the initial temperature and r is the ramping rate. Then $T(\delta, \tau) = T_{sat}(1 + \delta + \sigma\tau)$ where $\tau = |r|t/T_{sat}$, $\sigma = \operatorname{sgn}(r)$, and $\delta = (T_0 - T_{sat})/T_{sat}$. Here δ represents the relative difference between the initial temperature for the LPE process and the saturation temperature at which the liquid phase is prepared; this difference is referred to as undercooling.

The details of linearizing the interface conditions are discussed elsewhere [26, p. 27]. Roughly speaking, the results are that the GaAlAs crystal grows when $T < T_{sat}$ and decays when $T > T_{sat}$. Also the interface concentration for aluminum in the solid increases in time for $\sigma = 1$ and decreases for $\sigma = -1$.

Although the concentrations of the constituents in both phases can be determined, principal attention is given to the aluminum concentration in the solid. This concentration is of particular interest from the practical view of manufacturing devices. The goal is to form regions in the solid of high aluminum concentration sandwiched between regions having little or no aluminum. Ghez and Small [10] have speculated that an internal maximum in the aluminum concentration of the solid could be produced by a temperature ramp which caused the crystal to decay. Applying Lemma 2.3, however, one can see that this is not case. In particular, the presence of the exponential factor in (2.3) assures that no internal maximum is possible.

For GaAlAs, $D_{Al} \simeq D_{As} \simeq 10^{-5}$ cm^2/sec while reported values for D_s vary rather widely from 10^{-9} cm^2/sec to as low as 10^{-17} cm^2/sec [9]. Since in most LPE processes $D_s \leqq 10^{-12}$, the range for $\sqrt{\varepsilon}$ is approximately $10^{-6} \leqq \sqrt{\varepsilon} \leqq 10^{-4}$. Assume that $r =$ 10°C/min and that $T_0 = T_{sat} \simeq 850$°C (both of these assumptions are typical) [10]. For $\sqrt{\varepsilon} = 10^{-4}$, the condition that τ be larger than $\sqrt{\varepsilon}$ requires that t be greater than roughly 1 sec. Note that this restriction does not present a difficulty for the linearization of the interface conditions since $\tau \ll 1$ only implies that $t \ll 10^4$ sec. Now assume that the liquid phase is sufficiently thick so that the semi-infinite approximation is valid for both phases and that $u_s^e = 0.175$. Then $u_{Al}^{\infty} = 3.25 \times 10^{-3}$ and $u_{As}^{\infty} = 2.71 \times 10^{-2}$. If $u_s^{\infty} = 0.1$, then Fig. 3.1 illustrates the thermal decay solution for these parameter values when $t = 1$, 10, and 100 sec and shows how narrow the boundary layer becomes. Indeed since the figure spans only 100 A (20–50 atomic layers), the continuum model is breaking down, and one would expect no increase in the aluminum concentration in the interior of the solid for times much larger than 10 sec. For smaller values of D_s the boundary layer disappears even more rapidly. A temperature ramp of this sort

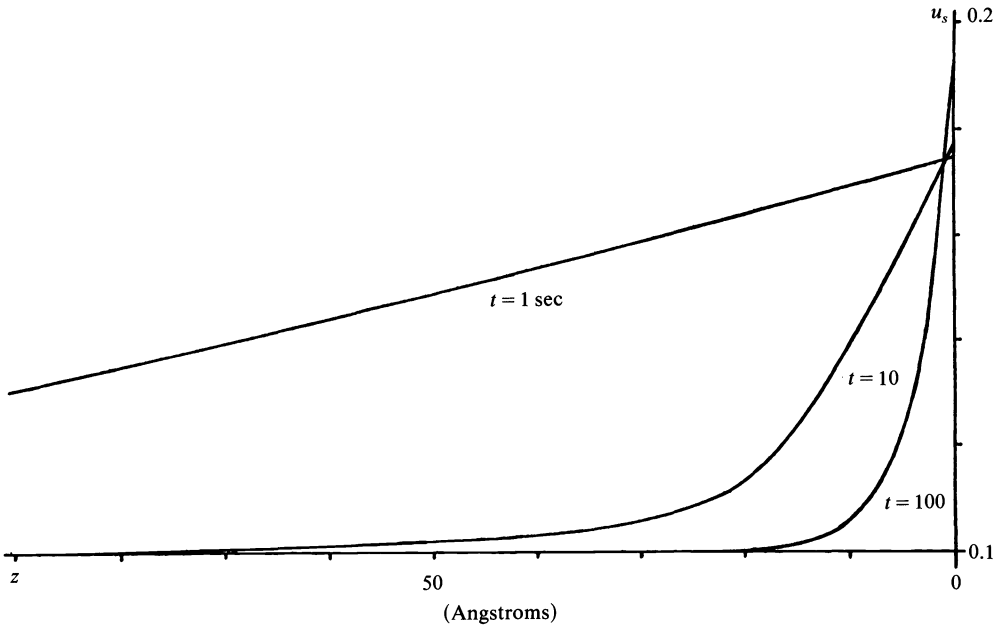


FIG. 3.1. *Graph of the long-time solution for temperature-ramp induced decay when* $T_0 = T_{sat} = 850°C$, $r = 10°C/min$, *and* $D_s = 10^{-13}$ cm^2/sec. *Note that* $z = 0$ *is the solid-liquid interface.*

would therefore not be useful in creating a shallow aluminum-rich region in an LPE crystal. From the solutions given in § 2, one sees that no other temperature ramp would create isolated aluminum-rich regions either.

Acknowledgments. My thanks go to David G. Schaeffer and Richard Ghez who have helped me greatly in this work.

REFERENCES

[1] L. I. RUBINSTEIN, *The Stefan Problem*, Trans. Math. Monographs, Amer. Math. Soc., Providence, RI, 1971.

[2] J. CRANK, *Free and Moving Boundary Problems*, Clarendon Press, Oxford, 1984.

[3] C. M. ELLIOTT AND J. R. OCKENDON, *Weak and Variational Methods for Moving Boundary Problems*, Research Notes in Math., 59, Pitman, Boston, 1982.

[4] A. FASANO AND M. PRIMICERIO, EDS., *Free Boundary Problems: Theory and Applications*, Vols. 1 and 2, Research Notes in Math., 78, 79, Pitman, Boston, 1983.

[5] ———, *General free-boundary problems for the heat equations*, J. Math. Anal. Appl., 57 (1977), pp. 694-723; 58 (1977), pp. 202-231; 59 (1977), pp. 1-14.

[6] A. B. TAYLER, *The mathematical formulation of Stefan problems*, in Moving Boundary Value Problems in Heat Flow and Diffusion, J. R. Ockendon and W. R. Hodgkins, eds., Clarendon Press, Oxford, 1974.

[7] L. I. RUBINSTEIN, *Free boundary problems*, Vols. I and II, Proceedings of a seminar held in Pavia (1979), E. Magenes, ed., Inst. Nazionale di Alta Matematica Francesco Severi, Rome, 1980.

[8] A. B. CROWLEY AND J. R. OCKENDON, *On the numerical solution of an alloy solidification problem*, Internat. J. Heat Mass Transfer, 22 (1979), pp. 941-947.

[9] M. B. SMALL AND R. GHEZ, *Growth and dissolution kinetics of* III-V *heterostructures formed by* LPE, J. Appl. Phys., 50 (1979), pp. 5322-5330.

[10] R. GHEZ AND M. B. SMALL, *Growth and dissolution kinetics of ternary* III-V *heterostructures formed by liquid-phase epitaxy.* III. *Effects of temperature programming*, J. Appl. Phys., 53 (1982), pp. 4907-4918.

[11] J. R. OCKENDON, *Techniques of analysis*, in Moving Boundary Value Problems in Heat Flow and Diffusion, J. R. Ockendon and W. R. Hodgkins, eds., Clarendon Press, Oxford, 1974.

[12] K. STEWARTSON AND R. T. WAECHTER, *On Stefan's problem for spheres*, Proc. Roy. Soc. London Ser. A, 348 (1976), pp. 415-426.

[13] A. B. TAYLER, *Mathematical Models in Applied Mechanics*, Clarendon Press, Oxford, 1986.

[14] W. W. MULLINS AND R. F. SEKERKA, *Morphological stability of a particle growing by diffusion or heat flow*, J. Appl. Phys., 34 (1963), pp. 323-329.

[15] ———, *Stability of a planar interface during solidification of a dilute binary alloy*, J. Appl. Phys., 35 (1964), pp. 444-451.

[16] J. S. LANGER, *Instabilities and pattern formation in crystal growth*, Rev. Modern Phys., 52 (1980), pp. 1-28.

[17] J. S. LANGER AND L. A. TURSKI, *Studies in the theory of interfacial stability*, Acta Metall., 25 (1977), pp. 1113-1119, 1121-1137.

[18] G. J. FIX, *Numerical methods for alloy solidification problems*, in Moving Boundary Problems, D. G. Wilson, A. D. Solomon and P. T. Boggs, eds., Academic Press, New York, 1978.

[19] A. BERMUDEZ AND C. SAGUEZ, *Mathematical formation and numerical solution of an alloy solidification problem*, in Free Boundary Problems: Theory and Application, Vols. 1 and 2, Research Notes in Math., 78, 79, A. Fasano and M. Primierio, eds., Pitman, Boston, 1983.

[20] R. GHEZ, *Expansions in time for the solution of one-dimensional Stefan problems of crystal growth*, Internat. J. Heat Transfer, 23 (1980), pp. 425-432.

[21] L. N. TAO, *The Stefan problem with arbitrary initial and boundary conditions*, Quart. Appl. Math., 36 (1978), pp. 223-233.

[22] I. M. GEL'FAND AND G. E. SHILOV, *Generalized Functions*, Vol. 1 (trans. by Eugene Saletan), Academic Press, New York, 1964.

[23] M. B. SMALL, R. GHEZ, R. M. POTEMSKI AND J. M. WOODALL, *The formation of* $Ga_{1-x}Al_xAs$ *layers on the surface of* GaAs *during continual dissolution into* Ga-Al-As *solutions*, Appl. Phys. Lett., 35 (1979), pp. 209-210.

[24] M. B. SMALL, R. GHEZ, R. M. POTEMSKI AND W. REUTER, *The dissolution kinetics of* GaAs *in undersaturated isothermal solutions in the* Ga-Al-As *system*, J. Electrochem. Soc., 127 (1980), pp. 1177-1182.

[25] M. B. SMALL, R. GHEZ, W. REUTER AND R. M. POTEMSKI, Al *diffusivity as a function of growth rate during the formation of* (GaAl)As *heterostructures by liquid phase epitaxy*, J. Appl. Phys., 52 (1981), pp. 814-817.

[26] J. D. FEHRIBACH, *Perturbation methods for solid diffusion in an infinite two phase Stefan problem: liquid-phase epitaxy in* GaAlAs, Ph.D. thesis, Duke University, Durham, NC, 1985.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

007

BIFURCATION ANALYSIS OF REACTION-DIFFUSION EQUATIONS VI. MULTIPLY PERIODIC TRAVELLING WAVES*

JAMES C. ALEXANDER† AND GILES AUCHMUTY‡

Abstract. The bifurcation of periodic travelling-wave solutions of a system of reaction-diffusion equations from a trivial solution is studied. We allow general periodicity conditions and show that the resulting bifurcation is characterized by the wave vector **k** and a wave speed c. Criteria for the global branching of such solutions are described and the results are applied to the Brusselator.

Key words. reaction-diffusion equations, global bifurcation, travelling waves, multiparameter bifurcation, multiply periodic

AMS(MOS) subject classification. 35D30, 35K20, 35B32

1. Introduction.

In recent years a lot of work has been done applying bifurcation theory to describe various special classes of solutions of reaction-diffusion equations. The bifurcation of stationary and time-dependent solutions is well-known. More recently considerable work has been done in describing wave-like solutions in problems with rotational symmetry; see (Alexander [1986]; Alexander and Auchmuty [1979]; Auchmuty [1979]; Auchmuty [1984]; Auchmuty and Nicolis [1976]). In this paper we shall study the bifurcation of travelling-wave solutions which are periodic in space.

Consider the general reaction-diffusion system

$$\frac{\partial u_i}{\partial t} = D_i \Delta u_i + f_i(u_1, \ldots, u_m; \mu), \tag{1.1}$$

on $R^n \times (0, \infty)$. Here $i = 1, \ldots, m$ and $u_i(\mathbf{x}, t)$ generally models the concentration of the ith chemical species at a point $(\mathbf{x}, t)$ of space-time. The Laplacian Δ is

$$\Delta v = \sum_{j=1}^{n} \frac{\partial^2 v}{\partial x_j^2}$$

and the functions $f_i: R^{m+1} \to R$, depending on a parameter μ, define the kinetics of the system.

Our interest is in studying the bifurcation of travelling-wave solutions of (1.1) which are also periodic in space. That is, we seek solutions of (1.1) of the form

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(x_1 - c_1 t, x_2 - c_2 t, \ldots, x_n - c_n t), \tag{1.2}$$

where $\mathbf{v}$ is also periodic on R^n. It may be periodic in each variable separately or more generally we treat the case where

* Received by the editors September 8, 1986; accepted for publication **November 5, 1986. This** work was supported in part by the National Science Foundation.

† Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742.

‡ Department of Mathematics, University of Houston, Houston, Texas 77004. The work of this author was supported in part by the Welch Foundation.

$$\mathbf{v}\left(\mathbf{x}+\sum_{j=1}^{n} k_j \mathbf{a}_j\right) = \mathbf{v}(\mathbf{x}) \tag{1.3}$$

for any n linearly independent vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ and any $\mathbf{k} = (k_1, k_2, \ldots, k_n) \in \mathbf{Z}^n$, the set of all n-tuples of integers.

The parameter μ will be regarded as the bifurcation parameter. In general, it need not occur in the kinetics but could arise elsewhere in the system, but this usually only causes minor changes in the analysis. It might appear that this is a bifurcation problem with $(n+1)$ parameters $(c_1, c_2, \ldots, c_n, \mu)$, but in fact there actually are only two parameters (c, μ), for bifurcation with a given wave vector $\mathbf{k} \in \mathbf{Z}^n$. This is described in section 4 where we give the general global bifurcation theorem. In sections 2 and 3 we formulate the problem mathematically as a fixed-point problem and then in section 5 we show how the analysis applies to the Brusselator in two and three space dimensions.

2. Description of the problem. Our interest is in studying the bifurcations of periodic travelling wave solutions of the reaction-diffusion system (1.1) from a trivial stationary solution. The pure reaction system may be written in vector form as

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}; \mu) \tag{2.1}$$

where $\mathbf{u}(t) = \big(u_1(t), u_2(t), \ldots, u_m(t)\big)^T$ and $\mathbf{f}(\mathbf{u}; \mu)$ has components $f_i(\mathbf{u}; \mu)$, $1 \le i \le m$. Here μ is a real parameter. We assume:

(F1) $\mathbf{f}: R^{m+1} \to R^m$ *is continuously differentiable and* $\mathbf{f}(\mathbf{0}; \mu) = \mathbf{0}$ *for all* $\mu \in R$.

Define $F(\mu) = \left(\frac{\partial f_i}{\partial u_j}(\mathbf{0}; \mu)\right)$ to be the Jacobian matrix of $\mathbf{f}(\,\cdot\,; \mu)$ at $\mathbf{u} = \mathbf{0}$. The assumption (F1) implies that $\mathbf{u}(\mathbf{x}, t) \equiv \mathbf{0}$ is a solution of (1.1) for all μ. It is called the *trivial solution* and we are interested in the bifurcation of certain wave-like solutions from it. For many particular systems, the trivial solution is not the zero solution, but a change of variables effects (F1).

Consider the problem of finding solutions of (1.1) of the form

$$\begin{aligned} \mathbf{u}(\mathbf{x}, t) &= \mathbf{v}(x_1 - c_1 t, x_2 - c_2 t, \ldots, x_n - c_n t) \\ &= \mathbf{v}(\mathbf{x} - \mathbf{c}t), \end{aligned} \tag{2.2}$$

where $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ is non-zero, and $\mathbf{v}$ is periodic of period α_j in the jth variable, $1 \le j \le n$. Let $\mathbf{e}_j$ be the unit vector in R^n whose jth coordinate is 1 and all other coordinates are 0, and let $\xi_j = x_j - c_j t$. Then for any $\mathbf{k} = (k_1, k_2, \ldots, k_n) \in \mathbf{Z}^n$, one has

$$\mathbf{v}\left(\boldsymbol{\xi} + \sum_{j=1}^{n} k_j \alpha_j \mathbf{e}_j\right) = \mathbf{v}(\boldsymbol{\xi}) \qquad \text{for all } \boldsymbol{\xi} \in R^n. \tag{2.3}$$

Moreover if $\mathbf{v}$ is a solution of (1.1), then its components v_i are solutions of the semi-linear elliptic system

$$D_i \Delta v_i + \sum_{j=1}^{n} c_j \frac{\partial v_i}{\partial \xi_j} + f_i(v_1, v_2, \ldots, v_m) = 0. \tag{2.4}$$

Here $1 \le i \le m$, $\boldsymbol{\xi} \in R^n$ and $\mathbf{v}$ obeys the periodicity conditions (2.3). One may regard (2.3), (2.4) as an elliptic system defined on the domain $\overline{\Omega}_\alpha = [0, \alpha_1] \times [0, \alpha_2] \times \cdots [0, \alpha_n]$

subject to the periodic boundary conditions

$$\begin{aligned} &\mathbf{v}(\boldsymbol{\xi}) = \mathbf{v}(\boldsymbol{\xi} + \alpha_j \mathbf{e}_j) \qquad \text{for all } \boldsymbol{\xi} \in \overline{\Omega}_\alpha \text{ with } \xi_j = 0, \\ &\frac{\partial \mathbf{v}}{\partial \xi_k}(\boldsymbol{\xi}) = \frac{\partial \mathbf{v}}{\partial \xi_k}(\boldsymbol{\xi} + \alpha_j \mathbf{e}_j), \qquad 1 \le j, k \le n. \end{aligned} \tag{2.5}$$

Thus the problem of finding periodic travelling-wave solutions of the form (2.2) is equivalent to solving the boundary-value problem (2.4)–(2.5).

Note that the vectors $\alpha_j \mathbf{e}_j$, $1 \le j \le n$, generate a lattice in R^n, a discrete subgroup of maximal rank n. The periodicity condition can be phrased: if $\mathbf{x}, \mathbf{x}' \in R^n$ are in the same coset of the lattice, the solutions are equal at $\mathbf{x}$ and $\mathbf{x}'$. The lattice of the $\alpha \mathbf{e}_j$, $1 \le j \le n$, is a rectangular lattice. It is possible to work with more general lattices; that is, more general periodicity conditions. Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ be n linearly independent points in R^n, and suppose $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{in})$. Let

$$\Omega_A = \left\{ \mathbf{x} \in R^n : \mathbf{x} = \sum_{j=1}^{n} \beta_j \mathbf{a}_j \text{ with } 0 \le \beta_j \le 1 \right\}.$$

Then Ω_A is an n-dimensional parallelepiped. When $n = 2$, it is a parallelogram with vertices $\mathbf{0}$, $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_1 + \mathbf{a}_2$. We look for solutions of (1.2) subject to the "skewed" periodicity conditions (1.3). For example, suppose $\mathbf{a}_1 = (1, 1)$ and $\mathbf{a}_2 = (0, 2)$. Then (1.3) reads

$$\mathbf{v}(x_1 + 1, x_2 + 1) = \mathbf{v}(x_1, x_2 + 2) = \mathbf{v}(x_1, x_2).$$

Such solutions are of course defined on the quotient space of R^n by the lattice. This space is a skewed flat n-dimensional torus. Topologically it is the product of n circles, one for each periodicity condition.

Let A be the (nonsingular) matrix (a_{ij}) and let $B = A^{-1}$. For the example just above

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 0 \end{pmatrix}, \qquad B = \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}.$$

Define $y_i = \sum_{j=1}^{n} b_{ij} x_j$, $1 \le i \le n$. Then $\mathbf{x} \in \overline{\Omega}_A$ if and only if $\mathbf{y} = B\mathbf{x} \in [0, 1]^n$, the cube of unit side in R^n. By the chain rule,

$$\frac{\partial u_i}{\partial x_j} = \sum_{k=1}^{n} b_{kj} \frac{\partial u_i}{\partial y_k},$$

so

$$\Delta_x u_i = \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} b_{kj} \frac{\partial^2 u_i}{\partial y_k \partial y_l} b_{lj} = \operatorname{tr} B^T D_y^2 u_i B = L_y u_i, \tag{2.6}$$

where $D_y^2 u_i = \left(\frac{\partial^2 u_i}{\partial y_k \partial y_l} \right)$ is the Hessian matrix of u_i with respect to y and $\operatorname{tr} A$ is the trace of A.

Now consider the problem of finding solutions of (1.1) of the form $\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(\mathbf{y} - \mathbf{c}t)$ with $\mathbf{y} = B\mathbf{x}$ as above. Then if $\mathbf{z} = \mathbf{y} - \mathbf{c}t$,

$$D_i L_z v_i + \sum_{j=1}^{n} c_j \frac{\partial v_i}{\partial z_j} + f_i(v_1, v_2, \ldots, v_m; \mu) = 0 \tag{2.7}$$

on R^n, subject to (2.3) with each $\alpha_j = 1$ and with L_z defined by (2.6). Alternately one may regard (2.7) as an elliptic system defined on the domain $\overline{\Omega}_1 = [0, 1] \times [0, 1] \times \cdots \times [0, 1]$ and subject to the periodicity conditions

(2.8)
$$\left.\begin{aligned} \mathbf{v}(\mathbf{z}) &= \mathbf{v}(\mathbf{z}+\mathbf{e}_j) \\ \frac{\partial \mathbf{v}}{\partial z_k}(\mathbf{z}) &= \frac{\partial \mathbf{v}}{\partial z_k}(\mathbf{z}+\mathbf{e}_j) \end{aligned}\right\} \quad \text{for all } \mathbf{z} \in \Gamma_j = \{\mathbf{z} \in \partial\Omega_1 : z_j = 0\}, \quad 1 \le j,k \le n.$$

Our interest is in finding nontrivial solutions of either (2.4)–(2.5) or (2.7)–(2.8). Note that (2.4)–(2.5) can be put in the form (2.7)–(2.8) by letting $z_j = \xi_j/\alpha_j$ for $1 \le j \le n$. Thus we concentrate on the latter problem.

3. Fixed-point formulation. The system of equations (2.7)–(2.8) may be formulated as a fixed-point problem. Consider the problem of solving the scalar equation

$$w - L_z w = g \qquad \text{on } \Omega_1 \tag{3.1}$$

subject to the boundary conditions

$$\left.\begin{aligned} w(\mathbf{z}) &= w(\mathbf{z}+\mathbf{e}_j) \\ \frac{\partial w}{\partial z_k}(\mathbf{z}) &= \frac{\partial w}{\partial z_k}(\mathbf{z}+\mathbf{e}_j) \end{aligned}\right\} \quad \text{for all } \mathbf{z} \in \Gamma_j, \quad 1 \le j \le n. \tag{3.2}$$

Let $L^2(\Omega_1)$ be the usual Lebesgue space of all square-integrable complex-valued measurable functions on Ω_1. Let $H_p^1(\Omega_1)$ (resp. $H_p^2(\Omega_1)$) be the usual Sobolev space of all complex-valued functions on Ω_1 which are restrictions of periodic functions on R^n, locally in H^1 (resp. H^2) and of period 1 in each variable. Suppose $w \in L^2(\Omega_1)$. Then w has a Fourier expansion

$$w(\mathbf{z}) = \sum_{\mathbf{k}\in\mathbf{Z}^n} w_{\mathbf{k}} \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle), \tag{3.3}$$

where $\langle \mathbf{k}, \mathbf{z}\rangle = \sum_{j=1}^n k_j z_j$ and $w_{\mathbf{k}} = \int_{\Omega_1} w(\mathbf{z}) \exp(-2\pi i\langle \mathbf{k}, \mathbf{z}\rangle)\, dz$. From Parseval's theorem, w is in $L^2(\Omega_1)$ if and only if $\sum_{\mathbf{k}\in\mathbf{Z}^n} |w_{\mathbf{k}}|^2 < \infty$. For $\mathbf{k} = (k_1, k_2, \ldots, k_n)$

$$\frac{\partial w}{\partial z_j}(\mathbf{z}) = 2\pi i \sum_{\mathbf{k}\in\mathbf{Z}^n} k_j w_{\mathbf{k}} \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle) \quad \text{and}$$

$$\frac{\partial^2 w}{\partial z_i \partial z_j}(\mathbf{z}) = -4\pi^2 \sum_{\mathbf{k}\in\mathbf{Z}^n} k_i k_j w_{\mathbf{k}} \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle).$$

Thus, using Parseval's theorem again,

$$w \in H_p^1(\Omega_1) \quad \text{if and only if} \quad \sum_{\mathbf{k}\in\mathbf{Z}^n} (1+|k|^2)|w_{\mathbf{k}}|^2 < \infty \quad \text{and}$$

$$w \in H_p^2(\Omega_1) \quad \text{if and only if} \quad \sum_{\mathbf{k}\in\mathbf{Z}^n} (1+|k|^2)^2|w_{\mathbf{k}}|^2 < \infty,$$

where $|\mathbf{k}|^2 = \sum_{j=1}^n k_j^2$.

THEOREM 1. *Suppose $g \in L^2(\Omega_1)$. Then there is a unique $w \in H_p^2(\Omega_1)$ obeying (3.1)–(3.2) and the mapping $G: L^2(\Omega_1) \to H_p^2(\Omega_1)$ defined by $Gg = w$ is a continuous linear map.*

Proof. Take inner products of (3.1) with w. Then

$$\int_{\Omega_1} |w(\mathbf{z})|^2\, dz - \langle L_z w, w\rangle = \int_{\Omega_1} \overline{w}(\mathbf{z}) g(\mathbf{z})\, dz.$$

Now

$$-\langle L_z w, w\rangle = -\int_{\Omega_A} \Delta w(\mathbf{x})\overline{w}(\mathbf{x})|\det B|\,dx = |\det B|\int_{\Omega_A} |\nabla w(\mathbf{x})|^2\,dx \geq 0$$

by the Gauss-Green theorem. Also $I - L_z: H_p^2(\Omega_1) \to L^2(\Omega_1)$ is a closed densely-defined linear operator on $L^2(\Omega_1)$, and this inequality implies it is positive-definite. From the Lax-Milgram theorem, it has a continuous inverse and the result follows from standard elliptic operator theory.

Let $\mathcal{P}^0$ be the space of all continuous real-valued functions $\mathbf{u}: \overline{\Omega}_1 \to R$ which satisfy $\mathbf{u}(\mathbf{z}) = \mathbf{u}(\mathbf{z} + \mathbf{e}_j)$ for $\mathbf{z} \in \Gamma_j$. Similarly let $\mathcal{P}^1$ be the space of all continuously differentiable real-valued functions on $\overline{\Omega}_1$ which satisfy (3.2). Note that $\mathcal{P}^0$ and $\mathcal{P}^1$ are Banach spaces under the norms

$$\|\mathbf{u}\|_0 = \sup_{\mathbf{z}\in\overline{\Omega}_1} |\mathbf{u}(\mathbf{z})|,$$

$$\|\mathbf{u}\|_1 = \sup_{\mathbf{z}\in\overline{\Omega}_1} \left[|\mathbf{u}(\mathbf{z})| + \sum_{k=1}^{n} \left|\frac{\partial u}{\partial z_k}(\mathbf{z})\right|\right].$$

COROLLARY. *Suppose $g \in \mathcal{P}^0$. Then the solution w of* (3.1)–(3.2) *is in $\mathcal{P}^1$ and the map $G: \mathcal{P}^0 \to \mathcal{P}^1$ is a compact linear map.*

Proof. If $g \in \mathcal{P}^0$, then $g \in L^p(\Omega_1)$ for all p, $1 < p < \infty$. Hence from standard elliptic theory, $w \in W^{2,p}(\Omega_1)$ and $G: L^p(\Omega_1) \to W^{2,p}(\Omega_1)$ is continuous. If $p > n$, then $W^{2,p}(\Omega_1)$ is compactly embedded in $\mathcal{P}^1$. This proves the result.

It is worth noting that

$$\begin{aligned} L_z \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle) &= -4\pi^2 \sum_{m=1}^{n}\sum_{l=1}^{n}\left(\sum_{j=1}^{n} b_{mj}b_{lj}\right) k_m k_l \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle) \\ &= -4\pi^2\langle \mathbf{k}B, \mathbf{k}B\rangle \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle) \end{aligned} \tag{3.4}$$

where $\langle\cdot,\cdot\rangle$ is the usual inner product on R^n. Thus if $g(\mathbf{z}) = \sum_{\mathbf{k}\in\mathbf{Z}^n} g_{\mathbf{k}} \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle)$, an explicit representation for the solution w of (3.1)–(3.2) is

$$w(\mathbf{z}) = \sum_{\mathbf{k}\in\mathbf{Z}^n} \frac{g_{\mathbf{k}} \exp(2\pi i\langle \mathbf{k}, \mathbf{z}\rangle)}{1 + 4\pi^2\langle \mathbf{k}B, \mathbf{k}B\rangle}. \tag{3.5}$$

Let $X_1 = \{(v_1, v_2, \ldots, v_m) : v_j \in \mathcal{P}^1, 1 \leq j \leq m\}$ and define $\|\mathbf{v}\|_1 = \sum_{j=1}^{m} \|v_j\|_1$. Define $\mathcal{G}(\cdot, \mathbf{c}, \mu): X_1 \to X_1$ by

$$\mathcal{G}_j(\mathbf{v}, \mathbf{c}, \mu) == D_j^{-1} G\left[\sum_{k=1}^{m} c_k \frac{\partial v_j}{\partial z_k} + f_j(\mathbf{v}; u) + D_j v_j\right]. \tag{3.6}$$

From the corollary to theorem 1, we see any classical wave solution $(\mathbf{v}, \mathbf{c})$ of (2.7)–(2.8) is a fixed point of $\mathcal{G}(\cdot, \cdot, \mu)$ in $X_1 \times R^n$. That is, $(\mathbf{v}, \mathbf{c})$ is a solution of

$$\mathbf{v} = \mathcal{G}(\mathbf{v}, \mathbf{c}, \mu) \tag{3.7}$$

PROPOSITION 3.1. *$\mathcal{G}: X_1 \to X_1$, defined by* (3.6), *is a compact map. It is Fréchet differentiable and*

$$\left[D\mathcal{G}(\mathbf{0}, \mathbf{c}, \mu)h\right]_j = D_j^{-1} G\left(\sum_{k=1}^{m} c_k \frac{\partial h_j}{\partial z_k} + D_j h_j + \sum_{l=1}^{m} \frac{\partial f_j}{\partial v_l}(\mathbf{0}; \mu) h_l\right). \tag{3.8}$$

The proof of this is identical to that of Proposition 2 in (Auchmuty [1979]).

Equation (3.5) is a fixed-point problem for a compact map of the Banach space X_1 to itself which depends on $n+1$ parameters $(c_1, c_2, \ldots, c_n, \mu)$. For all $(\mathbf{c}, \mu) \in R^{n+1}$, we have $\mathbf{v} = \mathbf{0}$ is a solution of (3.7), so we look for global branches of solutions of (3.7) bifurcating from this basic solution.

4. Bifurcation results. The formulation of the last section enables us to use general functional analytic methods for bifurcation theory. We use global results, as described in (Alexander [1978]) or (Alexander and Fitzpatrick [1979]); of course local results are subsumed.

First it is worth noting that the possible bifurcation points for (3.5) can be characterized in terms of eigenvalues of $m \times m$ matrices. Given a lattice of periodicity conditions, let B be the associated matrix. For $\mathbf{k} \in \mathbf{Z}^n$, let $\lambda_{\mathbf{k}} = 4\pi^2 \langle \mathbf{k}B, \mathbf{k}B \rangle$ be the associated eigenvalue of the Laplacian. Let D be the diagonal matrix with entries D_i, $1 \le i \le n$, let $F(\mu)$ be the Jacobian of $f(\,\cdot\,, \mu)$ at $u = 0$, and define the $m \times m$ matrix

$$F_{\mathbf{k}}(\mu) = F(\mu) - \lambda_{\mathbf{k}} D. \tag{4.1}$$

PROPOSITION 4.1. *If $(\mathbf{0}, \hat{\mathbf{c}}, \hat{\mu})$ is a bifurcation point for (3.7) (or the system (2.7)–(2.8)), then there exists $\mathbf{k} \in \mathbf{Z}^n$ such that $F_{\mathbf{k}}(\hat{\mu})$ has purely imaginary eigenvalues.*

Proof. If $(\mathbf{0}, \hat{\mathbf{c}}, \hat{\mu})$ is a bifurcation point for (3.7), then from the implicit function theorem, 1 is an eigenvalue of $D\mathcal{G}(\mathbf{0}, \hat{\mathbf{c}}, \hat{\mu})$. Let h be the corresponding eigenfunction. From (3.8), h is a solution of

$$D_j\left(L_z h_j\right) + \sum_{l=1}^{n} \hat{c}_l \frac{\partial h_j}{\partial z_l} + \sum_{l=1}^{m} F_{jl}(\hat{\mu}) h_l = 0, \tag{4.2}$$

and satisfies periodicity conditions like (2.8). Expand h in a Fourier series:

$$h(\mathbf{z}) = \sum_{\mathbf{k} \in \mathbf{Z}^n} \mathbf{a}_{\mathbf{k}} \exp(2\pi i \langle \mathbf{k}, \mathbf{z} \rangle),$$

with each $\mathbf{a}_{\mathbf{k}}$ in C^m. From (3.4), for some $\mathbf{k} \in \mathbf{Z}^n$,

$$-\lambda_{\mathbf{k}} D \mathbf{a}_{\mathbf{k}} + 2\pi i \langle \hat{\mathbf{c}}, \mathbf{k} \rangle \mathbf{a}_{\mathbf{k}} + F(\hat{\mu}) \mathbf{a}_{\mathbf{k}} = 0$$

for some nonzero $\mathbf{a}_{\mathbf{k}}$. Thus $\pm 2\pi i \langle \hat{\mathbf{c}}, \mathbf{k} \rangle$ must be eigenvalues of $F_{\mathbf{k}}(\mu)$.

As usual, the necessary condition is not sufficient for bifurcation. Standard crossing and non-degeneracy conditions are needed. The situation is still more complicated. There are ostensively $n+1$ parameters, the n components of $\mathbf{c}$ and μ. Moreover there are a considerable number of degeneracies. For example, suppose the periodicity lattice is given by the unit vectors $\mathbf{e}_j$. Then all interchanges of coordinates preserve the problem and virtually all the eigenvalues are highly degenerate. It is possible to analyze the problem using equivariant bifurcation theory; however for travelling waves, we can obtain more information by dealing directly with the problem. In particular, we decompose it into simple bifurcation problems, one for each $\mathbf{k}$. We see directly that the solutions we obtain are plane-wave solutions. Generically in f, these are all the wave solutions on primary bifurcation branches. Let

$$\Delta_{\mathbf{k}} = \left\{ \mathbf{v} \in \mathcal{P}^1 : \mathbf{v}(\mathbf{z}) = \mathbf{v}(\mathbf{z} + \mathbf{z}') \text{ whenever } \langle \mathbf{z}', \mathbf{k} \rangle = 0 \right\}.$$

If $\mathbf{k} \neq \mathbf{0}$, $\Delta_{\mathbf{k}}$ is a proper closed subspace of $\mathcal{P}^1$. If $\mathbf{v} \in \Delta_{\mathbf{k}}$, there is a function $\tilde{\mathbf{v}}: R \to R^n$ which is periodic of period $k_0 = \gcd\{k_i : k_i \neq 0, 1 \le i \le n\}$, such that

$\mathbf{v}(\mathbf{z}) = \tilde{\mathbf{v}}(\langle \mathbf{k}, \mathbf{z}\rangle)$ for all $\mathbf{z} \in R^n$. Thus if $\tilde{c} = \langle \mathbf{c}, \mathbf{k}\rangle$, define

$$\mathbf{v}(\mathbf{y} - \mathbf{c}t) = \tilde{\mathbf{v}}(\langle \mathbf{k}, \mathbf{y} - \mathbf{c}t\rangle) = \tilde{\mathbf{v}}(\langle \mathbf{k}, \mathbf{y}\rangle - \tilde{c}t).$$

That is $\mathbf{u}(\mathbf{x}, t) = \tilde{\mathbf{v}}(\langle B^T\mathbf{k}, \mathbf{x}\rangle - \tilde{c}t)$. Wave solutions of (3.1) in $\Delta_{\mathbf{k}}$ are actually functions only of the scalar variable $\langle B^T\mathbf{k}, \mathbf{x}\rangle - \tilde{c}t$.

Let

$$\mathcal{S}_{\mathbf{k}} = \left\{(\mathbf{v}, \tilde{c}, \mu) \in \Delta_{\mathbf{k}} \times R^2 : \mathbf{v} = \mathcal{G}(\mathbf{v}, \mathbf{c}, \mu) \text{ holds with } \tilde{c} = \langle \mathbf{c}, \mathbf{k}\rangle\right\},$$

and let

$$\mathcal{S}_{0,\mathbf{k}} = \left\{(\mathbf{0}, c, \mu) : (c, \mu) \in R^2\right\}.$$

Thus $\mathcal{S}_{0,\mathbf{k}}$ is the set of trivial solutions of (3.7) in $\Delta_{\mathbf{k}}$ and $\mathcal{S}_{\mathbf{k}}$ is the set of all solutions of (3.7) in $\Delta_{\mathbf{k}}$. Let $\mathcal{C}$ be a maximal connected subset of $\mathcal{S}_{\mathbf{k}} \setminus \mathcal{S}_{0,\mathbf{k}}$ with $(\mathbf{0}, \hat{c}, \hat{\mu}) \in \overline{\mathcal{C}}$. We say there is a *global branch of solutions* of (3.7) in $\Delta_{\mathbf{k}}$ bifurcating at $(\mathbf{0}, \hat{c}, \hat{\mu})$ provided one or more of the following holds:

(i) $\mathcal{C}$ is unbounded in $\Delta_{\mathbf{k}} \times (0, \infty) \times R$ or,

(ii) there exists $(\mathbf{v}, 0, \hat{\mu}) \in \overline{\mathcal{C}}$ or,

(iii) there exists $(\tilde{c}, \tilde{\mu}) \neq (\hat{c}, \hat{\mu})$ such that $(\mathbf{0}, \tilde{c}, \tilde{\mu}) \in \overline{\mathcal{C}}$.

Our main result is the following bifurcation theorem. For convenience of application, the bifurcation criteria are stated in terms of the matrices $F_{\mathbf{k}}(\mu)$. Fix $\mathbf{k}$.

THEOREM 2. *Suppose f satisfies* (F1) *and that $\mathcal{G}$, X, $F_{\mathbf{k}}(\mu)$, and $\Delta_{\mathbf{k}}$ are as above. Suppose there exists $\hat{\mu} \in R$ and $\mathbf{k} \neq \mathbf{0}$ such that*

(i) (*nonsingularity*) *$F_{\mathbf{k}}(\hat{\mu})$ is nonsingular,*

(ii) (*nonresonance*) *$F_{\mathbf{k}}(\hat{\mu})$ has a pair of simple eigenvalues $\pm i\nu$ and no other integer multiple of $\pm i\nu$ is an eigenvalue of $F_{\mathbf{k}}(\hat{\mu})$,*

(iii) (*transversality*) *the continuation of $i\nu$ in μ crosses the imaginary axis as μ crosses $\hat{\mu}$.*

Then there is a global branch of travelling-wave solutions of (3.7) *in $\Delta_{\mathbf{k}}$ bifurcating at $(\mathbf{0}, \nu/2\pi, \hat{\mu})$. Moreover the solutions $v_i(\mathbf{z})$ are classical solutions of* (2.7)-(2.8).

Proof. Both the linear equations (4.2) and the nonlinear equations (2.7) and (3.7) respect $\Delta_{\mathbf{k}}$, so we can restrict the analysis to $\Delta_{\mathbf{k}}$. If $\mathbf{v} \in \Delta_{\mathbf{k}}$, then

$$\mathbf{v}(\mathbf{z}) = \sum_{l=-\infty}^{\infty} \mathbf{v}_l \exp(2\pi i l\langle \mathbf{k}, \mathbf{z}\rangle) = \tilde{\mathbf{v}}(\langle \mathbf{k}, \mathbf{z}\rangle),$$

and thus conditions (i)—(iii) and Theorem 1 of (Alexander and Fitzpatrick [1979]) imply the existence of a global branch of travelling-wave solutions of (3.7) in $\Delta_{\mathbf{k}}$ bifurcating from $(\mathbf{0}, \hat{c}, \hat{\mu})$. From Proposition 4.1, $\nu = 2\pi\hat{\mathbf{c}}$. If $\mathbf{v} \in \Delta_{\mathbf{k}}$, then $\mathbf{v} \in \mathcal{P}^1$ and hence $L_z v_i$ is continuous on Ω_1 from (2.7). Hence each v_i is in $W^{2,\infty}(\Omega_1)$ from usual elliptic theory. This implies that $L_z v_i \in W^{1,\infty}(\Omega_1)$ for each i and hence v_i is a classical solution of (2.7)–(2.8) as required. The proof is complete.

Comments.

1. The bifurcation conditions (i)–(iii) are of standard type and are generically valid.
2. These travelling waves have wave-fronts given by

$$\langle B^T\mathbf{k}, \mathbf{x}\rangle - ct = \text{constant},$$

   so the fronts are hyperplanes in R^n. Moreover these waves are periodic functions of t of period k_0/c.
3. When $\mathbf{k} = \mathbf{0}$ and $F_0(\hat{\mu})$ is singular, bifurcating steady states occur, while if $\mathbf{k} = \mathbf{0}$ and the other conditions of the theorem hold, standing waves bifurcate via a

straightforward Hopf bifurcation.

4. Note that although we originally sought solutions involving a vector of wave speeds $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, it turns out that these bifurcating waves may be characterized by a scalar c. Thus although we originally had an $(n+1)$-parameter problem, the bifurcation analysis is a standard type involving only two parameters.
5. One can use the implicit function theorem to construct these waves near the bifurcation point and to first order they can be represented in terms of the real and imaginary parts of $\exp(2\pi i\langle \mathbf{k}, \mathbf{y} - \mathbf{c}t\rangle)$ multiplying the eigenvectors of $F_{\mathbf{k}}(\hat{\mu})$ associated with the eigenvalues $\pm i\nu$. More precisely, if the eigenvector of $F_{\mathbf{k}}(\hat{\mu})$ associated to $i\nu$ has components γ_i (defined up to a complex scalar), to first order the wave solutions have components

$$\epsilon|\gamma_1| \sin\left(\sum_{i=1}^{N} k_i y_i - \hat{\mathbf{c}}t + \arg\gamma_1\right), \ldots, \epsilon|\gamma_n| \sin\left(\sum_{i=1}^{N} k_i y_i - \hat{\mathbf{c}}t + \arg\gamma_n\right),$$

in terms of the $y_i = \sum_{j=1}^{n} b_{ij} x_j$, where ϵ is a parameter along the branch. These are harmonic in each component. The amplitude in the ith component is $\epsilon|\gamma_i|$ and the relative phases between components are the differences between the $\arg\gamma_i$.
6. Recall that the solutions could be defined on a quotient space of R^n, a torus, which is the product of n circles. The vector $\mathbf{k} = (k_1, \ldots, k_n)$ indexes a parallel family of circles on the torus which wind around k_i times in the ith direction. This family of circles is the direction of motion of the wave front. On perpendicular hyperplanes, the wave is constant. Indeed use of the function space $\Delta_{\mathbf{k}}$ reduces the analysis to bifurcation theory on the circle indexed by $\mathbf{k}$, which has radius $\sqrt{\sum_{i,j,l=1}^{N} b_{ji} b_{li} k_j k_l}$. In particular, local higher-order bifurcation analysis can be done, for example to determine the criticality of the bifurcating branch. The analysis here reduces to an analysis on a circle, such as in (Auchmuty and Nicolis [1976]). From the point of view of equivariant theory, the vectors $\mathbf{k}$ index the characters (irreducible representations) of the symmetry group $(S^1)^n$. Although there are other symmetries, these seem to be the ones relevant to travelling waves.
7. We have introduced the bifurcation parameter μ in the reaction term. This is the most common formulation. However, other parameters are also reasonable bifurcation parameters, in particular one or more of the diffusion coefficients d_i or even the size of the underlying torus (compare (Auchmuty [1982])).

5. Examples. Consider the problem of looking for periodic travelling-wave solutions of the Brusselators in two and three space dimensions. The Brusselator is a model chemical reaction whose analysis is described in (Nicolis and Prigogine [1977, part II]). It is often used as a test of machinery. The case of periodic one-dimensional waves was treated in (Auchmuty and Nicolis [1976]). The equations are

$$\begin{aligned} \frac{\partial X}{\partial t} &= D_1 \Delta X - (B+1)X + X^2 Y + A, \\ \frac{\partial Y}{\partial t} &= D_2 \Delta Y + BX - X^2 Y. \end{aligned} \tag{5.1}$$

For all A, B, $X(\mathbf{x}, t) \equiv A$, $Y(\mathbf{x}, t) \equiv B/A$ are solutions. Letting $u_1(\mathbf{x}, t) = X(\mathbf{x}, t) - A$, $u_2(\mathbf{x}, t) = Y(\mathbf{x}, t) - B/A$, we obtain

$$\begin{aligned} \frac{\partial u_1}{\partial t} &= D_1 \Delta u_1 + (B-1)u_1 + A^2 u_2 + h(u_1, u_2), \\ \frac{\partial u_2}{\partial t} &= D_2 \Delta u_2 - Bu_1 - A^2 u_2 - h(u_1, u_2), \end{aligned} \tag{5.2}$$

where h contains quadratic and cubic terms.

First we look for travelling-wave solutions of (5.2) in two space dimensions which are periodic of period 1 in x_1 and of period $l > 0$ in x_2. Let $\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(x_1 - c_1 t, x_2 - c_2 t) = \mathbf{v}(\xi_1, \xi_2)$. Then on R^2

$$\begin{aligned} D_1 \Delta v_1 + (c \cdot \nabla) v_1 + f_1(v_1, v_2) &= 0, \\ D_2 \Delta v_2 + (c \cdot \nabla) v_2 + f_2(v_1, v_2) &= 0, \end{aligned} \tag{5.3}$$

where $c \cdot \nabla = c_1 \frac{\partial}{\partial \xi_1} + c_2 \frac{\partial}{\partial \xi_2}$, $f_1(v_1, v_2) = (B-1)v_1 + A^2 v_1 + h(v_1, v_2)$, and $f_2(v_1, v_2) = -Bv_1 - A^2 v_2 - h(v_1, v_2)$. The periodicity conditions are:

$$\mathbf{v}(\xi_1 + 1, \xi_2) = \mathbf{v}(\xi_1, \xi_2) = \mathbf{v}(\xi_1, \xi_2 + l) \qquad \text{for all } (\xi_1, \xi_2) \in R^2. \tag{5.4}$$

Let $z_1 = \xi_1$, $z_2 = \xi_2 / l$ and $\mathbf{v}(\xi_1, \xi_2) = \mathbf{w}(\xi_1, \xi_2 / l)$; then $\mathbf{w}$ obeys (2.3) and

$$d_i L w_i + (c \cdot \nabla) w_i + f_i(\mathbf{w}, B) = 0, \tag{5.5}$$

where

$$L = \frac{\partial^2}{\partial^2 z_1^2} + \frac{1}{l^2} \frac{\partial^2}{\partial^2 z_2^2}.$$

The parameter B plays the role of μ so the matrices $F_{\mathbf{k}}(B)$ are

$$\begin{aligned} F_{\mathbf{k}}(B) &= \begin{pmatrix} B-1 & A^2 \\ -B & -A^2 \end{pmatrix} - 4\pi^2 \left(k_1^2 + \frac{k_2^2}{l^2} \right) \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \\ &= \begin{pmatrix} B - 1 - \lambda_{\mathbf{k}} D_1 & A^2 \\ -B & -A^2 - \lambda_{\mathbf{k}} D_2 \end{pmatrix}, \end{aligned}$$

with $\lambda_{\mathbf{k}} = 4\pi^2 \left(k_1^2 + k_2^2 l^{-2} \right)$.

In the next two theorems we need the expression for $\Delta(\lambda_{\mathbf{k}}) = \det F_{\mathbf{k}}(B)$ when $F_{\mathbf{k}}(B) = 0$. It is

$$\Delta(\lambda_{\mathbf{k}}) = A^2 + \lambda_{\mathbf{k}}(D_1 - D_2)A^2 - \lambda_{\mathbf{k}}^2 D_2^2.$$

THEOREM 3. *Suppose* $\mathbf{k} = (k_1, k_2) \in \mathbf{Z}^2 \backslash \{(0,0)\}$ *and* $\Delta(\lambda_{\mathbf{k}}) > 0$. *Then there is a global branch of periodic travelling-wave solutions of* (5.3) *obeying* (5.4) *bifurcating at* $B_{\mathbf{k}} = 1 + A^2 + \lambda_{\mathbf{k}}(D_1 + D_2)$ *with* $\hat{c} = (2\pi)^{-1} \sqrt{\Delta(\lambda_{\mathbf{k}})}$. *The solutions on this branch have the form*

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(k_1 x_1 + k_2 l^{-1} x_2 - ct), \tag{5.6}$$

with $\mathbf{v}$,c *varying along the branch and* $\mathbf{v}$ *being a vector-valued function of period* 1.

Proof. This results from verifying the conditions of Theorem 2. At $B_{\mathbf{k}}$, (i) and (ii) hold trivially and (iii) holds by computation. Hence the result follows.

In the three-dimensional case a very similar result is true. Suppose we seek travelling-wave solutions which have period 1 in x_1, l_2 in x_2 and l_3 in x_3 with l_2, $l_3 > 0$.

THEOREM 4. *Suppose* $\mathbf{k} = (k_1, k_2, k_3) \in \mathbf{Z}^3 \backslash \{(0,0,0)\}$, $\lambda_{\mathbf{k}} = k_1^2 + k_2^2 l_2^{-2} + k_3^2 l_3^{-2}$, *and* $\Delta(\lambda_{\mathbf{k}}) > 0$. *Then there is a global branch of periodic travelling-wave solutions of* (5.3) *bifurcating at*

$$B_{\mathbf{k}} = 1 + A^2 + 4\pi^2 \lambda_{\mathbf{k}}(D_1 + D_2)$$

with $\hat{c} = (2\pi)^{-1} \sqrt{\Delta(\lambda_{\mathbf{k}})}$. *The solutions on this branch have the form*

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{v}(k_1 x_1 + k_2 l_2^{-1} x_2 + k_3 l_3^{-1} x_3 - ct), \tag{5.7}$$

with $\mathbf{v},c$ *varying along the branch and* $\mathbf{v}$ *being a periodic vector-valued function of period* 1.

Proof. The only changes here are the new boundary conditions. The proof is the same as that for Theorem 3 and (5.7) is the analog of (5.6).

REFERENCES

J. C. ALEXANDER, *Bifurcation of zeroes of parametrized functions*, J. Funct. Anal., 29 (1978), pp. 37–53.

J. C. ALEXANDER, *Asymmetric travelling wave solutions of reaction-diffusion equations*, in Nonlinear Functional Analysis and its Applications, F. Browder, ed., AMS Proc. Symp. Pure Math., 45, American Mathematical Society, Providence, RI, 1986, pp. 7–16.

J. C. ALEXANDER AND J. F. G. AUCHMUTY, *Global bifurcation of waves*, Manuscripta Math., 27 (1979), pp. 159–166.

J. C. ALEXANDER AND P. M. FITZPATRICK, *The homotopy of certain spaces of nonlinear operators and its relation to global bifurcation of the fixed points of parametrized condensing operators*, J. Funct. Anal., 34 (1979), pp. 87–106.

J. F. G. AUCHMUTY, *Bifurcating waves*, in Bifurcation Theory and Applications in Scientific Disciplines, O. Gurel and O. E. Rössler, eds., Annals, 316, New York Acad. Sci., 1979, pp. 236–278.

J. F. G. AUCHMUTY, *Bifurcation analysis of reaction-diffusion equations,* IV. *Size dependence*, in Instabilities, Bifurcations and Fluctuations in Chemical Systems, L. E. Reidel and W. Schieve, eds., Univ. of Texas Press, Austin, TX, 1982, pp. 3–31.

J. F. G. AUCHMUTY, *Bifurcation analysis of reaction-diffusion equations,* V. *Rotating waves on a disc*, in Partial Differential Equations and Dynamical Systems, W. E. Fitzgibbon, ed., Research Notes in Mathematics, 101, Pitman, Marshfield, MA, 1984, pp. 173–181.

J. F. G. AUCHMUTY AND G. NICOLIS, *Bifurcation analysis of reaction-diffusion equations,* III. *Chemical oscillations*, Bull. Math. Biol., 38 (1976), pp. 325–350.

G. NICOLIS AND I. PRIGOGINE, *Self-Organization in Nonequilibrium Systems*, John Wiley & Sons, New York, 1977.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

008

QUASILINEAR EVOLUTION EQUATIONS IN NONCLASSICAL DIFFUSION*

KENNETH KUTTLER† AND ELIAS AIFANTIS‡

Abstract. After describing the motivation leading to some nonclassical diffusion equations, we formulate a general abstract nonlinear evolution equation and establish existence of solutions. Then we return to the original equation and discuss particular initial-boundary value problems.

Key words. existence, modeling

AMS(MOS) subject classifications. 35A05, 35G30, 35K60, 35K70

Introduction. A general framework based on the approach of continuum mechanics has been proposed recently by Aifantis [1] for a systematic development of diffusion models. In this method, the diffusing substance is viewed as a continuum subject to two kinds of forces: an internal body force vector arising from its interaction with the matrix and a stress tensor that the diffusing substance exerts on itself.

By introducing constitutive equations for the stress tensor and the internal force vector, we can obtain classes of diffusion behavior which take into account viscosity and higher-order gradient effects. Various diffusion models are thus generated within a unified mathematical framework.

For example, if the stress tensor is assumed to depend on the concentration and the gradient of the flux, while the internal body force is viewed as a drag proportional to the flux, a pseudoparabolic partial differential equation of the type studied by Ting [2] is obtained. This yields a physically realistic model of diffusion for situations where the effects of viscosity cannot be ignored. Similarly, the equation of spinodal decomposition of Cahn [3] can be obtained within this general formalism by including second gradients of the solute concentration in the constitutive equation for the stress tensor to allow for long-range effects. For a further discussion of the method and the development of many other examples, we refer to [1].

A central problem in the development of these new models is to determine which of the resulting partial differential equations are well posed. This is not always obvious, especially if nonlinear or time dependent equations are being considered. In a preceding paper [4], the questions of existence and uniqueness were resolved for a class of linear partial differential equations resulting when the stress $\underset{\sim}{T}$ is a linear function of the concentration, its gradients up to second order, and the gradient of the flux, while the internal body force $\underset{\sim}{f}$ is a linear function of the flux. The corresponding constitutive equations are thus

$$\begin{aligned} \underset{\sim}{T} &= c_1\rho\underset{\sim}{1} + c_2 \operatorname{tr}(\nabla \underset{\sim}{j})\underset{\sim}{1} + c_3 \operatorname{tr}(\nabla^2\rho)\underset{\sim}{1}, \\ \underset{\sim}{f} &= -\underset{\sim}{\alpha}(\underset{\sim}{x}, t)\underset{\sim}{j}, \end{aligned} \tag{0.1}$$

where $\underset{\sim}{j}$ is the flux, ρ is the concentration, $\nabla \underset{\sim}{j}$ denotes the first gradient of j ($\nabla \underset{\sim}{j} = j_{i,j}$), $\nabla^2\rho$ the second gradient of ρ ($\nabla^2\rho = \rho_{,ij}$), tr is the trace and $\underset{\sim}{\alpha}(\underset{\sim}{x}, t)$ is a nonsingular

* Received by the editors February 4, 1985; accepted for publication December 1, 1986. This work was supported by the Michigan Technological University creativity grants program, the SM program of the National Science Foundation and the MM program of Michigan Technological University.

† Department of Mathematical and Computer Sciences, Michigan Technological University, Houghton, Michigan 49931.

‡ Department of Mechanical Engineering and Engineering Mechanics, Michigan Technological University, Houghton, Michigan 49931.

symmetric matrix. The spatial and temporal dependence of $\underset{\sim}{\alpha}$ models the inhomogeneity of the interaction between the solid and the diffusing substance. The reason for neglecting nonhydrostatic components in the expression for the stress tensor in $(0, 1)_1$ is also discussed in [4].

Next we introduce (0.1) into the balance equations of mass and momentum which, on neglecting inertia forces, take the form

$$\begin{aligned} &\rho_t + \operatorname{div} \underset{\sim}{j} = 0, \\ &\operatorname{div} \underset{\sim}{T} + \underset{\sim}{f} = 0. \end{aligned} \tag{0.2}$$

This operation yields the following linear evolution equation whose existence and uniqueness have been studied earlier [4].

$$\frac{\partial}{\partial t}(\rho - c_2 \operatorname{div} (\underset{\sim}{\alpha}^{-1}\nabla\rho)) = -c_1 \operatorname{div} (\underset{\sim}{\alpha}^{-1}\nabla\rho) - c_3 \operatorname{div} (\underset{\sim}{\alpha}^{-1}\nabla(\Delta\rho)). \tag{0.3}$$

In the present paper we consider a more general physical situation by allowing the constants c_1 and c_3 in $(0.1)_1$ to be functions of ρ. Roughly speaking, this means physically that we consider situations where the diffusion coefficient is concentration dependent. In this connection, our results are most suitable for problems in the nonlinear theory of spinodal decomposition, where c_2 vanishes identically. Thus our present expression for the stress $\underset{\sim}{T}$ takes the form

$$\underset{\sim}{T} = c_1(\rho)\underset{\sim}{1} + c_2 \operatorname{tr} (\nabla \underset{\sim}{j})\underset{\sim}{1} + c_3(\rho) \operatorname{tr} (\nabla^2\rho)\underset{\sim}{1}. \tag{0.4}$$

On substituting (0.4) in the balance laws (0.2), we find that ρ satisfies a partial differential equation of the form

$$\rho_t - c_2 \operatorname{div} (\underset{\sim}{\beta}(x, t)\nabla\rho_t) + \operatorname{div} (\underset{\sim}{\beta}\nabla(c_3(\rho)\Delta\rho)) + \operatorname{div} (\underset{\sim}{\beta}\nabla c_1(\rho)) = 0 \tag{0.5}$$

where $\underset{\sim}{\beta}(\underset{\sim}{x}, t) = \underset{\sim}{\alpha}^{-1}(\underset{\sim}{x}, t)$. As explained in [4], we could also have allowed c_1, c_2 and c_3 to depend on $\underset{\sim}{x}$ and t, but in any case, we would have arrived at an equation of the following general form:

$$\begin{aligned} &\frac{\partial}{\partial t}\Big(\rho - \sum_{ij} \partial_i(\tilde{D}_{ij}(\underset{\sim}{x}, t)\partial_j\rho)\Big) - \sum_{ij} \partial_i(D_{ij}(\underset{\sim}{x}, t; \rho)\partial_j\rho) \\ &\qquad + \sum_{|\alpha|,|\beta|\leqq 2} (-1)^{|\beta|} D^\beta(E_{\alpha\beta}(\underset{\sim}{x}, t; \rho)D^\alpha\rho) = h(\underset{\sim}{x}, t), \end{aligned} \tag{0.6}$$

where α, β are multi-indices [5], h is a source function and the rest of the coefficient functions are to be specified later. Since this extra generality does not create essential difficulties in the mathematical treatment, we will consider this last equation. We show that weak solutions to (0.6) exist by formulating a corresponding abstract problem and obtaining estimates that allow the use of a fixed point theorem.

The plan of the paper is as follows: Section 1 is a review of the linear version of these equations. Section 2 contains an abstract result, motivated by (0.6), which might be useful in other problems in the theory of nonlinear evolution equations. This result is used in § 3 to reduce the question of existence of weak solutions for (0.6) to the verification of a coercivity inequality. In § 4 we actually prove that this inequality holds. This way, we obtain the existence of solutions for (0.6) to some specific initial-boundary value problems. We use the standard notation for Banach spaces: If V is a Banach space, V' denotes its dual, $\to$ denotes strong convergence, and $\rightharpoonup$ denotes weak convergence. If i: $V \to W$ is an injection map, i^* denotes the dual map defined by

$$\langle i^* f, u\rangle = \langle f, iu\rangle = \langle f, u\rangle.$$

1. The linear equations. Before dealing with (0.6), we discuss briefly the main results of the earlier paper [4] in which D_{ij} and $E_{\alpha\beta}$ do not depend on ρ. In this earlier paper, the assumptions were

$$\begin{aligned} &\tilde{D}_{ij} = \tilde{D}_{ji}, \\ &\tilde{D}_{ij} \text{ is bounded, measurable and } C^1 \text{ in } t, \\ &\sum_{ij} \tilde{D}_{ij}\xi_i\xi_j \geqq 0 \quad \text{for } \xi = (\xi_1, \xi_2, \xi_3) \in R^3, \\ &D_{ij} \in L^\infty(\Omega \times [0, T]), \\ &E_{\alpha\beta} \in L^\infty(\Omega \times [0, T]), \end{aligned} \tag{1.1}$$

along with a coercivity inequality similar to (2.5) of the present paper.

Under these assumptions, existence, uniqueness and continuous dependence results were obtained for a large class of initial-boundary value problems associated with the linear version of (0.6). The existence part was based on the verification of this coercivity inequality which allowed the use of the main existence theorem of [6] or [11]. The uniqueness may be obtained as a special case of the uniqueness theorem of [6].

To be more specific, sufficient conditions were given for well-posedness of weak solutions of the following initial boundary value problems:

$$\frac{\partial}{\partial t}\left(\rho - \sum_{ij} \partial_i(\tilde{D}_{ij}(\underset{\sim}{x}, t)\partial_j\rho)\right) - \sum_{ij} \partial_i(D_{ij}(\underset{\sim}{x}, t)\partial_j\rho) + \Delta^2\rho = g(\underset{\sim}{x}, t), \tag{1.2.1}$$

$$\rho(\underset{\sim}{x}, 0) = \rho_0(\underset{\sim}{x}), \tag{1.2.2}$$

along with either the boundary conditions

$$\frac{\partial\rho}{\partial n}(\underset{\sim}{x}, t) = \frac{\partial w}{\partial n}(\underset{\sim}{x}, t), \qquad x \in \partial\Omega, \tag{1.2.3}$$

$$\frac{\partial}{\partial t}\left(\sum_{ij} \tilde{D}_{ij}\partial_j\rho n_i\right) + \sum_{ij} (D_{ij}\partial_i\rho)n_i - \frac{\partial(\Delta\rho)}{\partial n} = l, \qquad x \in \partial\Omega, \tag{1.2.4}$$

or the boundary conditions

$$\rho(\underset{\sim}{x}, t) = c(t) + w(\underset{\sim}{x}, t), \qquad x \in \partial\Omega, \tag{1.2.5}$$

$$\int_{\partial\Omega} \frac{\partial\rho}{\partial t}(\underset{\sim}{x}, t)\, ds = \int_{\partial\Omega} \frac{\partial w}{\partial n}(\underset{\sim}{x}, t)\, ds, \tag{1.2.6}$$

$$\int_{\partial\Omega} \left[\frac{\partial}{\partial t}\left(\sum_{ij} \tilde{D}_{ij}\partial_j\rho\right)n_i + \sum_{ij} D_{ij}\partial_j\rho n_i - \frac{\partial(\Delta\rho)}{\partial n}\right] ds = \int_{\partial\Omega} l(\underset{\sim}{x}, t)\, ds, \tag{1.2.7}$$

$$\Delta\rho(x, t) - k(x, t) = r(t) \tag{1.2.8}$$

where in (1.2.3)-(1.2.8), $r(t), c(t)$ are unknown functions and w and l are given functions.

In the present paper, we shall use the existence and uniqueness of solutions to an appropriate abstract version of the linear problem, along with a well-known generalization of the Brouer fixed point theorem, to establish the existence of solutions to initial-boundary value problems corresponding to (0.6). We shall show that, just as in the linear case, the verification of an appropriate inequality is sufficient.

2. The abstract equation. For the sake of both generality and simplicity in the presentation, we obtain existence of solutions to (0.6) as a special case of an abstract result. We introduce the following hypotheses and conventions:

(2.1) V, W are reflexive Banach spaces $V \subseteq W$, $\|v\|_V \geqq \|v\|_W$, so that $V \subseteq W \subseteq i^* W' \subseteq V'$.

On defining $B(t)$ as a continuous linear map from W to W', we will assume

$$\begin{aligned} &\langle B(t)u, u\rangle \geqq 0, \\ &\langle B(t)u, v\rangle = \langle B(t)v, u\rangle, \\ &t \to B(t)u \quad \text{is in } C^1(0, T; W'). \end{aligned} \tag{2.2}$$

We will also make use of the space

$$\begin{aligned} X = \{u \in L^2(0, T; V) \text{ such that} \\ (Bu)' \in L^2(0, T; V')\}\ \|u\|_X = \|u\|_{L^2(0,T;V)} + \|(Bu)'\|_{L^2(0,T;V')} \end{aligned} \tag{2.3}$$

where by $(Bu)'$ we mean a unique function in $L^2(0, T; V')$, such that

$$\int_0^T (Bu)'(t)\phi(t)\, dt = -\int_0^T i^* B(t)u(t)\phi'(t)\, dt \quad \text{for all } \phi \in C_0^\infty(0, T).$$

It follows that X is a reflexive Banach space.

For each $w \in L^2(0, T; V)$, let $A(w)$ be a continuous linear map from $L^2(0, T; V)$ to $L^2(0, T; V')$ satisfying the following property:

(2.4)
(i) $\sup\{\|A(w)\|,\ w \in L^2(0, T; V)\} = Q < \infty$,

(ii) If $u_n \rightharpoonup u$ in X and $v_n \rightharpoonup v$ in $L^2(0, T; V)$ then for some subsequence $u_{n'}$, $v_{n'}$, $A(u_{n'})v_{n'} \rightharpoonup A(u)v$ in $L^2(0, T; V')$.

Moreover, by introducing the definitions
(i) $B: L^2(0, T; W) \to L^2(0, T; W')$ is given by $B(t)u(t) = Bu(t)$,
(ii) $B': L^2(0, T; W) \to L^2(0, T; W')$ is given by $B'(t)u(t) = B'u(t)$,
we can postulate the following coercivity inequality:

$$2\langle A(w)u, u\rangle + \lambda\langle Bu, u\rangle + \langle B'u, u\rangle \geqq C_1 \|u\|^2_{L^2(0,T;V)}, \tag{2.5}$$

for some $\lambda \in \mathbb{R}$ independent of w and $C_1 > 0$.

Finally, for each $u \in L^2(0, T; V)$, let $f(u) \in L^2(0, T; V')$ satisfy the following properties:

(2.6)
(i) $\sup\{\|f(u)\|_{L^2(0,T;V')},\ u \in X\} = P < \infty$,

(ii) If $u_n \rightharpoonup u$ in X, then $f(u_{n'}) \rightharpoonup f(u)$ in $L^2(0, T; V')$, for some subsequence $u_{n'}$.

With these assumptions, we can state the main existence theorem.

THEOREM 1. *With* (2.1)-(2.6) *valid and* $u_0 \in V$, *there exists* $u \in X$ *such that*

$$\begin{aligned} &(Bu)' + A(u)u = f(u), \\ &i^* Bu(0) = i^* B(0)u_0. \end{aligned} \tag{2.7}$$

Proof. It follows from (2.5) and [6] that for each $w \in X$ there exists a unique solution $u \in X$ to the problem

$$\begin{aligned} &(Bu)' + A(w)(u) = f(w), \\ &i^* Bu(0) = i^* B(0) u_0. \end{aligned} \tag{2.8}$$

On denoting this solution by $\psi(w)$ we have $\psi : X \to X$.

Next we make use of three lemmas whose proof may be found in [6].

LEMMA 1. *For each $u \in X$,*

$$\langle (Bu)'(t), u(t)\rangle = \frac{1}{2}\left[\frac{d}{dt}\langle Bu(t), u(t)\rangle + \langle B'(t)u(t), u(t)\rangle\right] \text{ a.e.} \tag{2.9}$$

Moreover $\langle Bu(t), u(t)\rangle$ is equal to an absolutely continuous function a.e. and point evaluation of $i^* Bu(\cdot)$ is a continuous map from X to V'.

LEMMA 2. *For $u, v \in X$, $\langle Bu(t), v(t)\rangle$ equals an absolutely continuous function a.e. denoted by $\langle Bu, v\rangle(\cdot)$. Moreover, there exists a constant M such that*

$$|\langle Bu, v\rangle(t)| \leqq M \|u\|_X \|v\|_X \quad \text{for all } t \in [0, T]. \tag{2.10}$$

LEMMA 3. *If $i^* Bv(0) = 0$ for $v \in X$, then there exists a sequence $\{v_n\} \subseteq X$ such that $\|v - v_n\|_X \to 0$ and $v_n(t) = 0$ in some neighborhood of 0.*

As a consequence of Lemmas 1-3 we can establish the following results:

$$\begin{aligned} &\text{(i)} \quad \langle Bu, u\rangle(t) - \langle Bu, u\rangle(0) + \int_0^t \langle B'(s)u(s), u(s)\rangle \, ds \\ &\qquad + 2\int_0^t \langle A(w)u(s), u(s)\rangle \, ds = 2\int_0^t \langle f(w)(s), u(s)\rangle \, ds, \end{aligned} \tag{2.11}$$

$$\text{(ii)} \quad \langle Bu, u\rangle(0) = \langle B(0)u_0, u_0\rangle. \tag{2.12}$$

Relation (2.11) is obtained by multiplying $(2.8)_1$ by u, using Lemma 1 and integrating the result from 0 to t. Relation (2.12) is derived by first using Lemma 3 to obtain a sequence $\{u_n\} \subseteq X$ with $u_n(t) = u_0$ near 0 and converging to u in X, and then using Lemma 2 together with the inequality

$$\begin{aligned} |\langle Bu, u\rangle(0) - \langle B(0)u_0, u_0\rangle| &= |\langle Bu, u\rangle(0) - \langle Bu_n, u_n\rangle(0)| \\ &\leqq |\langle B(u_n - u), u_n\rangle(0)| + |\langle Bu, u_n - u\rangle(0)| \\ &\leqq M(\|u\|_X + \|u_n\|_X)\|u_n - u\|_X. \end{aligned} \tag{2.13}$$

With (2.11) and (2.12) valid and the use of (2.5) and (2.6), we can establish the following main inequality:

$$\begin{aligned} &\langle Bu, u\rangle(t) + C_1 \|u\|^2_{L^2(0,t;V)} \\ &\qquad \leqq \lambda \int_0^t \langle Bu(s), u(s)\rangle \, ds + 2P\|u\|_{L^2(0,t;V)} + \langle B(0)u_0, u_0\rangle. \end{aligned} \tag{2.14}$$

On subtracting $C_1\|u\|^2_{L^2(0,t;V)}$ from both sides, we first note that

$$2P\|u\|_{L^2(0,t;V)} - C_1\|u\|^2_{L^2(0,t;V)} \leqq \frac{P^2}{C_1}. \tag{2.15}$$

It then follows that

$$\langle Bu, u\rangle(t) \leqq \left(\langle B(0)u_0, u_0\rangle + \frac{P^2}{C_1}\right) e^{\lambda t} \tag{2.16}$$

by an application of Gronwall's inequality. Having thus established that $\langle Bu, u\rangle(t)$ is bounded uniformly for $t\in[0, T]$ independently of w, (2.14) implies that $\|\psi w\|_{L^2(0,T;V)}$ is bounded independently of w. It now follows from $(2.8)_1$, (2.4) and (2.6) that $\|\psi w\|_X$ is bounded independently of w. Moreover, if $N\geqq\sup\{\|\psi w\|_X, w\in X\}$ and $S=\{w\in X$ such that $\|w\|_X\leqq N\}$, it follows that $\psi: X\to S$.

As a final step in the proof of the theorem we establish the following.

LEMMA 4. *$\psi: X\to X$ is weakly continuous.*

Proof. Let $u_n\rightharpoonup u$ in X. If ψu_n fails to converge weakly to ψu, then by selecting a subsequence also denoted by u_n, we may assume $u_n\rightharpoonup u$ in X and $\psi u_n\rightharpoonup z\neq\psi u$ in X. By utilizing the definition of ψ

$$\begin{aligned}&(B(\psi u_n))'+A(u_n)\psi u_n=f(u_n),\\ &i^*B(\psi u_n)(0)=i^*B(0)u_0,\end{aligned}\tag{2.17}$$

and properties (2.4) and (2.6) together with Lemma 1, we obtain

$$\begin{aligned}&(Bz)'+A(u)z=f(u),\\ &i^*Bz(0)=i^*B(0)u_0.\end{aligned}\tag{2.18}$$

Obviously, (2.18) contradicts the assumption that $z\neq\psi u$; therefore ψ is weakly continuous and Lemma 4 is established.

Thus, the proof of Theorem 1 is now completed by invoking Tykhanov's fixed point theorem [7] which asserts that ψ has a fixed point in S.

3. The nonlinear partial differential equation. Here we apply the abstract result of § 2 to the question of existence of solutions for initial-boundary value problems associated with the generalized diffusion equation (0.6). We will assume the following general properties for the relevant coefficients:

$$\begin{aligned}&\text{(i)}\quad \tilde{D}_{ij}=\tilde{D}_{ji},\\ &\text{(ii)}\quad \tilde{D}_{ij}\text{ is bounded, measurable, and } C^1\text{ in } t,\\ &\text{(iii)}\quad \sum_{ij}\tilde{D}_{ij}\xi_i\xi_j\geqq 0\quad\text{for } \underset{\sim}{\xi}=(\xi_1,\xi_2,\xi_3)\in\mathbb{R}^3,\\ &\text{(iv)}\quad \sup\left\{\sum_{ij}|D_{ij}(t,x;r)|+\sum_{|\alpha|,|\beta|\leqq 2}|E_{\alpha\beta}(x,t;r)|,\ (x,t,r)\in\bar{\Omega}\times[0,T]\times\mathbb{R}\right\}<\infty,\\ &\text{(v)}\quad r\to D_{ij}(x,t;r)\text{ and }r\to E_{\alpha\beta}(x,t;r)\text{ are continuous and real valued}\end{aligned}\tag{3.1}$$

where Ω is a bounded open set in $\mathbb{R}^3$.

With these, and in order to cast (0.6) in the abstract form of (2.7), we let V be a closed subspace of $H^2(\Omega)$, $W=H^1(\Omega)$, $H=L^2(\Omega)$, and for $y\in L^2(0, T; H^2(\Omega))$ we let $\tilde{A}(y)$ be a continuous linear map from $L^2(0, T; V)$ to $L^2(0, T; V')$ defined by

$$\begin{aligned}\langle\tilde{A}(y)u, v\rangle=&\int_0^T\int_\Omega D_{ij}(x,t;y(t)(x))\partial_i u(t)(x)\partial_j v(t)(x)\,dx\,dt\\ &+\int_0^T\int_\Omega E_{\alpha\beta}(x,t;y(t)(x))D^\alpha u(t)(x)D^\beta v(t)(x)\,dx\,dt,\end{aligned}\tag{3.2}$$

and $B(t)$; $W\to W'$ defined by

$$\langle B(t)u, v\rangle=\int_\Omega u(x)v(x)+\tilde{D}_{ij}(x,t)\partial_i u(x)\partial_j v(x)\,dx,\tag{3.3}$$

where summation over repeated indices is assumed. With $B(t)$ given by (3.3), it is clear that (2.2) holds.

Having already specified definitions and hypotheses (2.1)-(2.3), we proceed by considering properties (2.4)-(2.6). Of these properties, (2.4) and (2.6) are verified in this section while the coercivity inequality (2.5) is examined in the next section. To do this we first prove the following lemma which is a generalization of a well-known result in [8, p. 57].

LEMMA 5. *If* $u_n \rightharpoonup 0$ *in* X, *then*

(i) $\lim_{n\to\infty} i^*Bu_n(t) = 0$ *in* V' *for each* $t \in [0, T)$,

(ii) $\lim_{n\to\infty} \langle Bu_n, u_n\rangle = 0$ *in* $L^1(0, T)$,

(iii) $\lim_{n\to\infty} u_n = 0$ *in* $L^2(0, T; H)$,

where each limit in (i)-(iii) *refers to the strong topology of the space indicated.*

Proof. We first note that $i^*Bu(\cdot)$ is an absolutely continuous function with values in V' since i^*Bu_n and $(i^*Bu_n)'$ are both in $L^2(0, T; V')$. Thus $i^*Bu_n(t)$ is well defined and

$$\begin{aligned} i^*Bu(t) &= -\frac{1}{s}\int_t^{t+s} (Bu_n)'(r)(t+s-r)\,dr + \frac{i^*}{s}\int_t^{t+s} Bu_n(r)\,dr \\ &= U_n + i^*V_n. \end{aligned} \tag{3.4}$$

Thus, for a given $\varepsilon > 0$, it follows that $\|U_n\|_{V'} \leqq \varepsilon$ for all n if s is small enough. With this choice for s and $w \in W$, we have

$$|\langle V_n, w\rangle_{W',W}| = \left|\int_0^T \left\langle B(r)\frac{1}{s}\chi_{[t,t+s]}(r)w, u_n(r)\right\rangle_{W'W} dr\right|. \tag{3.5}$$

Since $B(\cdot)(1/s)\chi_{[t,t+s]}(\cdot) \in L^2(0, T; W')$ and $u_n \rightharpoonup 0$ in $L^2(0, T; V)$, the right-hand side of (3.5) converges to 0. But $w \in W$ was arbitrary and therefore $V_n \rightharpoonup 0$ in W'. The inclusion map of V into W is compact and thus i^*V_n converges strongly to 0 in V'. This proves (i) since $\varepsilon > 0$ was arbitrary.

To prove (ii), let $\varepsilon > 0$ be given. If α is large enough, we have

$$\begin{aligned} \int_0^T \langle Bu_n(t), u_n(t)\rangle\,dt &\leqq \frac{\alpha^2}{2}\int_0^T \|i^*Bu_n(t)\|_{V'}^2\,dt + \frac{1}{2\alpha^2}\int_0^T \|u_n(t)\|_V^2\,dt \\ &\leqq \varepsilon + \frac{\alpha^2}{2}\int_0^T \|i^*Bu_n(t)\|_{V'}^2\,dt. \end{aligned} \tag{3.6}$$

In view of Lemma 1, the term $\|i^*Bu_n(t)\|_{V'}^2$ is bounded independently of t and n. Therefore, the Dominated Convergence Theorem [9] and (i) imply the convergence to zero of the last term of (3.6), and since ε was arbitrary, part (ii) follows. Part (iii) is clearly implied by (ii). This completes the proof of the lemma.

As a final step in establishing the validity of (2.4) and (2.6), we introduce the definitions

$$\begin{aligned} &\text{(i)} \quad A(v) = \tilde{A}(v+w), \\ &\text{(ii)} \quad f(v) = -\tilde{A}(w+v)w - i^*B'w - i^*Bw' + g, \end{aligned} \tag{3.7}$$

where w and w' are both in $L^2(0, T; H^2(\Omega))$, $v \in L^2(0, T; V)$, $g \in L^2(0, T; V')$ and i is the injection map of V into $H^2(\Omega)$. Then the following lemma can be established.

LEMMA 6. *Hypotheses* (2.4) *and* (2.6) *hold.*

Proof. By $(3.1)_4$, it is clear that there exists $Q < \infty$ such that $\|A(u)\| \leqq Q$ for all $u \in L^2(0, T; V)$. Now let $u_n \rightharpoonup u$ in X and let $v_n \rightharpoonup v$ in $L^2(0, T; V)$. From Lemma 5,

$\lim_{n\to\infty} \|u_n - u\|_{L^2(0,T;H)} = 0$. Therefore a subsequence of $\{u_n\}$ converges to u a.e. in t and x. Then (2.4) follows from the Dominated Convergence Theorem and $(3.1)_5$. Hypothesis (2.6) also holds by similar arguments.

In view of the above arguments, we have reduced the problem of existence of solutions to the abstract evolution equation (2.7) in the special context of § 3 to the verification of the coercivity inequality (2.5). This will be discussed in the next section. For the convenience of presentation, however, this inequality will be assumed to hold in the remaining part of this section in order to provide the explicit form of the partial differential equation that we are concerned with here.

To do this, we define $g \in L^2(0, T; V')$ to be given by the relation

$$\begin{aligned}\langle g, v\rangle = \int_0^T \Big[\int_\Omega h(t)(x)v(t)(x)\, dx + \int_{\partial\Omega} l(t)(x)v(t)(x)\, dA \\ + \int_{\partial\Omega} k(t)(x) \frac{\partial v(t)}{\partial n}(x)\, dA \Big]\, dt\end{aligned} \tag{3.8}$$

where $h \in L^2(0, T; H)$, $(k, l) \in L^2(0, T; L^2(\partial\Omega))$ and $\partial\Omega$ is assumed to be a smooth two-dimensional manifold. Since the trace map from $H^1(\Omega)$ to $L^2(\partial\Omega)$ is continuous, it is clear that g is in $L^2(0, T; V')$. On assuming that (2.5) holds, it follows that Theorem 1 implies the existence of $u \in X$ satisfying the equation

$$\begin{aligned}&-\int_0^T \int_\Omega [u(t)(x)v(x) + \tilde{D}_{ij}(x, t)\partial_i u(t)(x)\partial_j v(x)]\, dx\, \phi'(t)\, dt \\ &\quad + \int_0^T \int_\Omega D_{ij}(x, t; w(t)(x) + u(t)(x))\partial_i u(t)(x)\partial_j v(t)(x)\, dx\, \phi(t)\, dt \\ &\quad + \int_0^T \int_\Omega E_{\alpha\beta}(x, t; w(t)(x) + u(t)(x))D^\alpha u(t)(x) D^\beta v(x)\, dx\, \phi(t)\, dt \\ &= \int_0^T \int_\Omega h(t)(x)v(x)\, dx\, \phi(t)\, dt \\ &\quad + \int_0^T \int_{\partial\Omega} \left(l(t)(x)v(x) + k(t)(x) \frac{\partial v(x)}{\partial n} \right) dA\, \phi(t)\, dt \\ &\quad - \int_0^T \int_\Omega D_{ij}(x, t; w(t)(x) + u(t)(x))\partial_i w(t)(x)\partial_j v(x)\, dx\, \phi(t)\, dt \\ &\quad - \int_0^T \int_\Omega E_{\alpha\beta}(x, t; w(t)(x) + u(t)(x))D^\alpha w(t)(x) D^\alpha v(x)\, dx\, \phi(t)\, dt \\ &\quad + \int_0^T \int_\Omega [w(t)(x)v(x) + \tilde{D}_{ij}(x, t)\partial_i w(t)(x)\partial_j v(x)]\, dx\, \phi'(t)\, dt,\end{aligned} \tag{3.9}$$

together with the initial condition

$$i^* Bu(0) = i^* B(0)(u_0 - w(0)) \tag{3.10}$$

for all $v \in V$ and $\phi \in C_0^\infty(0, T)$ provided $u_0 - w(0) \in V$.

On restricting v to be in $C_0^\infty(\Omega)$ and letting $z = u + w$ we see that a measurable representative of z is a weak solution of the partial differential equation (0.6) subject to the initial condition $i^* Bz(0) = i^* B(0)u_0$. Stable boundary conditions are obtained by properly selecting the space V, while variational boundary conditions are obtained

by the use of the divergence theorem in (3.9). This leads to the formulation of a variety of initial-boundary value problems, representative examples of which are considered in the next section.

4. Boundary value problems. In this section we consider particular initial-boundary value problems pertaining to (0.6) and establish existence of weak solutions by utilizing the results derived earlier. As mentioned previously, our task has been reduced to the verification of the coercivity inequality (2.5). Here this is accomplished in relation to specific forms of the associated boundary conditions. Three different sets of such conditions are considered below. The first set corresponds to Dirichlet type and coercivity is obtained as a result of Garding's inequality. The other two examples include variational-type boundary conditions and coercivity is established by other means.

4.1. Dirichlet boundary conditions. We choose $V = H_0^2(\Omega)$ and assume that the coefficients $E_{\alpha\beta}$ for $|\alpha| = |\beta| = 2$ are independent of ρ and are continuous on $\bar{\Omega} \times [0, T]$. Moreover, we suppose that they obey the strong ellipticity condition

$$\sum_{|\alpha|,|\beta|=2} E_{\alpha\beta}(x, t)\underset{\sim}{\xi}^{\alpha}\underset{\sim}{\xi}^{\beta} \geqq C|\underset{\sim}{\xi}|^4 \quad \text{for all } \underset{\sim}{\xi} \in \mathbb{R}^3, \tag{4.1}$$

so that the conditions of Garding's inequality [5] are satisfied. It then follows that (2.5) holds.

Thus, we have obtained existence of a weak solution to (1.6), denoted by z, along with boundary and initial conditions of the form

$$\begin{aligned} &z(t) - w(t) \in H_0^2(\Omega) \text{ a.e.,} \\ &i^* Bz(0) = i^* B(0)u_0, \end{aligned} \tag{4.2}$$

where i is the injection map of V into W. In less abstract fashion, the boundary condition $(4.2)_1$ can be expressed as

$$\begin{aligned} &z(t, x) = w(t, x), \qquad x \in \partial\Omega, \\ &\partial_i z(t, x) = \partial_i w(t, x), \qquad x \in \partial\Omega, \end{aligned} \tag{4.3}$$

where w is the prescribed function defined earlier. Roughly speaking, (4.3) suggests that in contrast to second-order problems, both the function and its derivatives need to be specified on the boundary for this class of fourth-order problems. These problems may be viewed as pertinent to the later stages of the important metallurgical process of spinodal decomposition, where nonlinear effects dominate.

4.2. Variational boundary conditions. In discussing boundary conditions of variational type, we consider a simplified form of the diffusion equation (0.6) as follows:

$$\frac{\partial}{\partial t}(\rho - \Delta\rho) - \partial_i(D(\rho)\partial_i\rho) + \Delta^2\rho = h. \tag{4.4}$$

This corresponds to assuming that the stress coefficient $c_3(\rho)$ in (1.4) is a constant and the mobility coefficient $\underset{\sim}{\alpha}$ in $(1.1)_2$ is a scalar α. Physically, these assumptions mean that nonlinear effects are retained in the dependence of the usual diffusion coefficient D but not in the small correcting terms, due to viscosity and surface tension.

We let $u_0 \in V \equiv \{u \in H^2(\Omega)$ such that $\partial u/\partial n = 0$ on $\partial\Omega\}$, with $\partial\Omega$ smooth. Then by the well-known theorem on elliptic regularity [10], $I - \Delta$ is a one-to-one and onto mapping from V to $L^2(\Omega)$. It follows $(I - \Delta)^{-1}$ is continuous by the open mapping theorem [9]. Therefore, there exists a $K > 0$ such that the following inequality holds:

$$\|u\|_V = \|(I - \Delta)^{-1}(I - \Delta)u\|_V \leqq K\|(I - \Delta)u\|_{L^2(\Omega)} \leqq K(\|u\|_{L^2(\Omega)} + \|\Delta u\|_{L^2(\Omega)}). \tag{4.5}$$

Then $\langle A(v)u, y\rangle = \langle \tilde{A}(v+w)u, y\rangle$ is of the form

$$\int_0^T \int_\Omega D(v(t)(x)+w(t)(x))\partial_i u(t)(x)\partial_j y(t)(x)\, dx\, dt + \int_0^T \int_\Omega \Delta u(t)(x)\Delta y(t)(x)\, dx\, dt \tag{4.6}$$

where $w(\cdot)$ and $w'(\cdot)$ are both in $L^2(0, T; H^2(\Omega))$ and $D(\cdot)$ is bounded and continuous. Similarly, $\langle B(t)u, v\rangle$ is of the form

$$\langle B(t)u, v\rangle = \int_\Omega (uv + \nabla u \cdot \nabla v)\, dx. \tag{4.7}$$

As a result of (4.5), it is easy to see through (4.6) and (4.7) that (2.5) holds. This establishes existence of solutions to (3.9) specialized to the present context. Applying then the divergence theorem, we obtain the existence of $u \in X$ such that $z = u + w$ satisfies (4.4) and the integral condition

$$\int_{\partial\Omega} \frac{\partial}{\partial n}(z_t)v + D(z)\frac{\partial z}{\partial n}v + \Delta z\frac{\partial v}{\partial n} - v\frac{\partial(\Delta z)}{\partial n}\, dA = \int_{\partial\Omega}\left(lv + k\frac{\partial v}{\partial n}\right) dA \tag{4.8}$$

for almost all values of t and for all $v \in V$.

Therefore z solves

$$\frac{\partial}{\partial t}(z - \Delta z) - \partial_i(D(z)\partial_i z) + \Delta^2 z = h, \tag{4.9}$$

along with the initial condition

$$\lim_{t\to 0+} \int_\Omega (z(t) - u_0)v + \nabla(z(t) - u_0)\cdot\nabla v\, dx = 0 \quad \text{for all } v \in V \tag{4.10}$$

and the boundary conditions

$$\begin{aligned} &\frac{\partial z(t, x)}{\partial n} = \frac{\partial w(t, x)}{\partial n} \quad \text{a.e. } t \text{ and } x \in \partial\Omega, \\ &\frac{\partial z_t(t, x)}{\partial n} + D(z(t, x))\frac{\partial z(t, x)}{\partial n} - \frac{\partial(\Delta z(t, x))}{\partial n} = l(t, x) \quad \text{a.e. } t \text{ and } x \in \partial\Omega \end{aligned} \tag{4.11}$$

where $(4.11)_1$ is stable resulting from the choice of V and $(4.11)_2$ is of a variational type resulting from the divergence theorem.

The initial condition (4.10) can be expressed in a more conventional form by noting that for $u \in X$, $Bu(t)$ is a function in $C(0, T; W')$. It follows that $u \in C(0, T; W)$ and thus $z = (u + w) \in C(0, T; W)$. Therefore, the limit in (4.10) can be taken inside the integral giving

$$\int_\Omega (z(0) - u_0)v + \nabla(z(0) - u_0)\cdot\nabla v\, dx = 0 \quad \text{for all } v \in V. \tag{4.12}$$

If $z(0) - u_0 \in V$, it follows that the initial condition (4.12) takes the usual form,

$$z(0, \cdot) = u_0(\cdot). \tag{4.13}$$

The condition that $z(0, \cdot) - u_0(\cdot) \in V$ is equivalent to saying that $(\partial z/\partial n)(0, \cdot) = (\partial w/\partial n)(0, \cdot) = \partial u_0(\cdot)/\partial n$ on $\partial\Omega$; that is, the initial condition $u_0(\cdot)$ and the boundary condition at $t = 0$, $w(0, \cdot)$ are compatible.

Next, we turn to a second example pertaining again to (4.4) but we now let $V=\{u\in H^2(\Omega)$ such that $u(x)=0$ on $\partial\Omega\}$. By reasoning similar to that of the previous example, (2.5) is again satisfied. Thus, in this case, we obtain the existence of a weak solution to the problem

$$
\begin{aligned}
&\frac{\partial}{\partial t}(z-\Delta z)-\partial_i(D(z)\partial_i z)+\Delta^2 z=h,\\
&z(t,\cdot)=w(t,\cdot)\quad \text{on } \partial\Omega \text{ for a.e. } t,\\
&\Delta z(t,\cdot)=k(t,\cdot)\quad \text{on } \partial\Omega \text{ for a.e. } t,\\
&\lim_{t\to 0+}\int_\Omega (z(t)-u_0)v+\nabla(z(t)-u_0)\cdot\nabla v\,dx=0,\qquad v\in V.
\end{aligned}
\tag{4.14}
$$

As before, $z(\cdot)$ is in $C(0, T; W)$ and if $z(0)-u_0\in V$, the initial condition $(4.14)_4$ takes the usual form $z(0,\cdot)=u_0(\cdot)$. In this case, the condition that $z(0)-u_0\in V$ is equivalent to the requirement that $u_0(\cdot)=w(0,\cdot)$ on $\partial\Omega$.

Other examples could be considered in a similar manner. Questions of existence of solutions to (0.6) or its specializations may thus be resolved by considering the verification of (2.5); that is the coercivity of a family of bilinear forms. This question of coercivity has been extensively studied and we refer to [10] for further discussion.

REFERENCES

[1] E. C. AIFANTIS, *On the problem of diffusion in solids*, Acta Mech., 37 (1980), pp. 265-296.

[2] T. W. TING, *Parabolic and pseudoparabolic partial differential equations*, J. Math. Soc. Japan, 21 (1969), pp. 440-453.

[3] J. W. CAHN, *On spinodal decomposition*, Acta Metallurgica, 9 (1979), pp. 795-901.

[4] K. L. KUTTLER AND E. C. AIFANTIS, *Existence and uniqueness in nonclassical diffusion*, Quart. Appl. Math., to appear.

[5] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[6] K. L. KUTTLER, *Time dependent implicit evolution equations*, Nonlinear Anal., 10 (1986), pp. 447-463.

[7] D. R. SMART, *Fixed Point Theorems*, Cambridge Univ. Press, 1974. (A generalization of the required fixed point theorem is stated on p. 32.)

[8] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non-lineares*, Dunod, Paris, 1969.

[9] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1974.

[10] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, New York, 1975.

[11] R. E. SHOWALTER, *Degenerate evolution equations and applications*, Indiana Univ. Math. J., 23 (1974), pp. 655-677.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

009

A THEOREM OF LA SALLE–LYAPUNOV TYPE FOR PARABOLIC SYSTEMS*

RAY REDHEFFER†, REINHARD REDLINGER‡ AND WOLFGANG WALTER‡

Abstract. This paper deals with the boundary value problem for a nonlinear system of parabolic differential equations for $u = u(t, x)$

$$u_t = Lu + f(u) \quad \text{in } \Omega, \qquad u(0, x) \text{ given}, \qquad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega$$

under the assumption that a Lyapunov function $V(z)$ for the corresponding ordinary differential equation system $u' = f(t, u)$ exists. In the case where L is one and the same selfadjoint elliptic operator of second order for all components of u, the real-valued function $U(t, x) = V(u(t, x))$ satisfies a parabolic differential inequality

$$U_t \leqq LU - c|u_x|^2 \qquad (c > 0).$$

It follows that u exists globally and is bounded if $u(0, x)$ is bounded. The limit set Λ^+ (as $t \to \infty$) of any solution u is nonempty and compact, it consists of constant functions only, it is an invariant set for $u' = f(u)$, and $\dot{V} = V_z \cdot f$ vanishes on Λ^+ (analogue of La Salle's stability theorem for ordinary differential equations). The results are then extended to quasilinear systems where $Lu = (a_{ij}(x, u)u_{x_j})_{x_j}$. In the case where different elliptic operators are involved, $u_t^k = L^k u^k + f^k(u)$ $(k = 1, \cdots, n)$, it is assumed that $a_{ij}^k(x) = c^k(x)a_{ij}(x)$ with $c^k > 0$. A Lyapunov functional $U(t) = \int_\Omega V(u(t, x))\, dx$ is employed, but the boundedness of solutions has to be assumed or obtained by other means.

Key words. parabolic systems, Lyapunov function, asymptotic behavior, limit set

AMS(MOS) subject classification. 35K40

1. Introduction.

We begin by briefly reviewing the ODE case. Let

$$u'(t) = f(u(t)) \tag{1}$$

be an autonomous system of n ordinary differential equations in an open set $D \subset R^n$, where $f: D \to R^n$ is continuous and such that the initial-value problems for (1) are uniquely solvable. This holds, for example, if f is locally Lipschitz continuous. Here a C^1 function $V: D \to R$ such that the set $D_c = \{z \in D: V(z) \leqq c\}$ is a compact subset of D for every $c \in V(D)$ is called a *Lyapunov function* for (1). The function

$$\dot{V}(z) = V_z \cdot f := \sum_{i=1}^{n} f_i \frac{\partial V}{\partial z_i},$$

the derivative of V in the direction f, has the property that

$$\frac{d}{dt} V(u(t)) = \dot{V}(u(t))$$

for any solution u of (1). The latter expression is called "the derivative of V along a solution" and is denoted by $\dot{V}(u)$.

Let u be a solution of (1) existing for $t \geqq 0$. The set $\Lambda^+ = \Lambda^+(u)$ defined by

$$\Lambda^+(u) = \{z \in R^n: z = \lim u(t_k) \text{ for some sequence } t_k \to \infty\}$$

* Received by the editors August 12, 1985; accepted for publication February 28, 1987.

† Department of Mathematics, University of California, Los Angeles, California 90024. Present address, Mathematisches Institut I, Universitat Karlsruhe, Kaiserstr. 12, D-7500 Karlsruhe 1, West Germany, under the auspices of the U.S. Special Program, Alexander von Hulmboldt-Stiftung.

‡ Mathematisches Institut I, Universitat Karlsruhe, Kaiserstr. 12, D-7500 Karlsruhe 1, West Germany.

is called the positive limit set of u. Let $\dot{V}(z) \leqq 0$ for $z \in D$. Then each solution of (1) exists for all $t \geqq 0$, the set Λ^+ is a compact nonempty subset of D, and dist $(u(t), \Lambda^+) \to 0$ as $t \to \infty$. It follows from this latter statement that Λ^+ is not only contained in the set

$$W = \{z \in D: \dot{V}(z) = 0\}$$

but is contained in the largest subset of this set which is invariant under (1). The description of Λ^+ is usually referred to as *La Salle's stability theorem.* Although various ingredients for La Salle's theorem were available previously, starting with the pioneering work of Lyapunov, it appears that the complete result was first formulated by La Salle (cf. [11] and the references there given).

The primary subject of this paper is the parabolic system

$$u_t^k = Lu^k + f^k(u) \quad \text{in } J \times \Omega \quad (k = 1, 2, \cdots, n). \tag{2}$$

In short $u_t = Lu + f(u)$, where $u = (u^1, u^2, \cdots, u^n) = u(t, x)$, $x = (x_1, x_2, \cdots, x_m)$, J is an interval $(0, T]$ or $(0, \infty)$, Ω is a bounded open set in R^m with smooth boundary, and L is a selfadjoint elliptic operator with coefficients independent of t,

$$L\phi = \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial \phi}{\partial x_j}\right). \tag{3}$$

Note that we have the same elliptic part in all equations, so that L can be regarded as operating on the real-valued function ϕ or as operating componentwise on the vector-valued function u. For the most part we shall deal with the Neumann problem, in which the initial and boundary conditions are

$$u(0, x) = \bar{u}(x) \quad \text{in } \bar{\Omega}, \qquad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } J \times \Gamma \tag{4}$$

where $\Gamma = \partial\Omega$ and $\partial/\partial\nu$ denotes the outer conormal derivative; that is,

$$\nu_i = \sum_{j=1}^{m} a_{ji} n_j,$$

where $n = (n_i)$ is the outer normal at $x \in \Gamma$. Our basic result, given explicitly in Theorems 1 and 2, can be roughly described as follows. If a convex Lyapunov function satisfying $\dot{V}(z) \leqq 0$ exists, then the parabolic case reduces to the ODE case. More precisely, solutions to (2), (4) exist for all $t \geqq 0$ and are bounded in $C^{2+\gamma}(\bar{\Omega})$ for some $\gamma > 0$. The limit set Λ^+ of any solution u, which is now a set in the space $C^{2+\gamma}$, consists of constant functions only. Considered as a subset of R^n, the limit set Λ^+ is an invariant set for the ODE (1) and the function $\dot{V}(z)$ vanishes on Λ^+. Hence Λ^+ is contained in the largest invariant subset (with respect to (1)) of the subset W defined above. It will be seen that

$$\text{dist}\,(u(t, \cdot), \Lambda^+) \to 0 \quad \text{as } t \to \infty,$$

where the distance is taken in the sense of the $C^{2+\gamma}$ norm. Hence, the largest invariant subset of W can be used to assess the asymptotic behavior not only for the ODE (1) but also for the PDEs (2) and (4).

As to method, our work relies heavily on the theory of parabolic differential inequalities. In the ODE case the function $U(t) = V(u(t))$ satisfies $U'(t) \leqq 0$ and hence is decreasing. The results of Lyapunov and their extension due to La Salle follow easily from this fact. In the parabolic case the corresponding real-valued function $U(t, x) = V(u(t, x))$ satisfies a parabolic inequality of the form

$$U_t \leqq LU - \alpha\beta|u_x|^2; \tag{5}$$

this fact is crucial for our method. Here α and β are positive constants such that $(a_{ij}) \geqq \alpha I_m$, $V_{zz} \geqq \beta I_n$, where I_k is the identity matrix of order k, V_{zz} is the Hessian of V, and the inequalities are interpreted in the sense of quadratic forms. The assumption $V_{zz} > 0$ is an additional requirement that is not needed in the ODE case.

The general problem considered here has a number of special cases that are of considerable interest in applications. Discussion of such cases can be found in the work of Alikakos [1]–[3], Chaffee [7], [8], Fleming [9], Hadeler [10], Leung [12], Mottoni, Orlandi and Tesei [13], Rothe [17], Webb [20] and in our own analysis of the prey-predator case [14]. These references make use of Lyapunov-type arguments, as we do, and [2] also indicates the importance of a condition of strict inequality such as $V_{zz} > 0$. We show by an example that, in fact, the weak inequality $V_{zz} \geqq 0$ is not sufficient.

2. Notation, assumptions and auxiliary theorems. Our basic notation is the same as in [14] but is briefly described for the reader's convenience. The Euclidean norm in R^m or R^n is denoted by $|\cdot|$. For a function $w: \bar{\Omega} \to R^n$ we set

$$\|w\|_0 = \sup |w(x)|, \quad [w]_\gamma = \sup \frac{|w(x) - w(y)|}{|x - y|^\gamma} \quad (x \neq y),$$

where $x \in \bar{\Omega}$, $y \in \bar{\Omega}$ and $0 < \gamma < 1$. Furthermore, $\|w\|_\gamma = \|w\|_0 + [w]_\gamma$. The spaces $C^\gamma(\bar{\Omega})$ or $C^{2+\gamma}(\bar{\Omega})$ consist of all functions with a finite norm $\|w\|_\gamma$ or

$$\|w\|_{2+\gamma} = \|w\|_0 + \sum_{i=1}^{m} \|D_i w\|_0 + \sum_{i,j=1}^{m} \|D_i D_j w\|_\gamma \qquad \left(D_i = \frac{\partial}{\partial x_i}\right),$$

respectively. The space $C^{1+\gamma}(\bar{\Omega})$ is defined similarly. Whenever a distinction between scalar- and vector-valued functions is not self-evident, we use notation such as $C^\gamma(\bar{\Omega}, R)$ or $C^{2+\gamma}(\bar{\Omega}, R^n)$. The gradient and Hessian of V are denoted by V_z and V_{zz}, respectively, and we set, by definition,

$$|u_x|^2 = \sum_{i=1}^{m} \sum_{k=1}^{n} |D_i u^k|^2.$$

We use a summation convention for repeated indices, the range of summation being from 1 to m or from 1 to n, as will be clear from the context. Thus, $L\phi = D_i(a_{ij} D_j \phi)$ and if $A = (a_{ij})$, $V_{zz} = (b_{ij})$, the inequalities $A \geqq \alpha I_m$, $V \geqq \beta I_n$ mean

$$a_{ij} \xi^i \xi^j \geqq \alpha |\xi|^2, \qquad b_{ij} \xi^i \xi^j \geqq \beta |\xi|^2$$

for $\xi \in R^m$ and for $\xi \in R^n$, respectively. The range of summation in the first case is from 1 to m and in the second, from 1 to n.

Concerning the domain Ω, the elliptic operator L, the function f and the initial value $\bar{u}$, we require the following:

(R1) Ω is a bounded, open, connected set in R^m with orientable boundary $\partial\Omega = \Gamma$ of class $C^{2+\gamma}$.

(R2) $D \subset R^n$ is open and $f: D \to R^n$ is locally Lipschitz continuous.

(R3) The matrix $A(x) = (a_{ij}(x))$ is of class $C^{1+\gamma}(\bar{\Omega})$ and $A \geqq \alpha I_m$ with $\alpha > 0$ constant.

(R4) $\bar{u}(x) \in C^{2+\gamma}(\bar{\Omega})$, $\partial \bar{u}/\partial \nu = 0$ on Γ, and $\bar{u}(x) \in D$ for $x \in \bar{\Omega}$.

If $u(t, x)$ is such that u, u_t, u_x and u_{xx} are continuous in $J \times \bar{\Omega}$, we write $u \in C^*(J)$. The following local existence theorem holds under the assumptions (R), and any other assumptions under which it holds would suffice for our purposes.

THEOREM 0. *Under the assumptions* (R) *the problem*

$$u_t = Lu + f(u), \qquad u(0, x) = \bar{u}(x), \qquad \frac{\partial u}{\partial \nu} = 0 \tag{6}$$

has a unique local solution $u \in C^*(J)$, $J = (0, T]$, $T > 0$. *The matrix* u_{xx} *is Hölder continuous in* x *and* $u(t, \cdot)$ *maps* $\bar{J}$ *continuously into* $C^{2+\gamma}(\bar{\Omega})$. *If an a priori estimate* $|u(t, x)| \leqq K$ *for* $0 \leqq t \leqq T$, $x \in \bar{\Omega}$ *can be established, where* K *is independent of* T, *then the solution exists for all* $t \geqq 0$ *and* $\sup_{t \geqq 0} \|u(t, \cdot)\|_{2+\gamma} < \infty$.

In (6) it is understood that the first equality holds in $J \times \Omega$ with $J = (0, \infty)$, the second in $\bar{\Omega}$, and the third in $J \times \Gamma$. To avoid unnecessary clutter, a similar convention is followed below.

The assertion of existence in Theorem 0 has been known for some time but prior to the investigation in [15], as far as we know, the global bound for $\|u\|_{2+\gamma}$ has been available only in special cases.

When $\|u\|_{2+\gamma}$ is bounded, the "path" $\{u(t, \cdot): t \geqq 0\}$ is a relatively compact subset of $C^2(\bar{\Omega})$. That is, for any sequence $t_k \to \infty$ there is a subsequence t_k^* such that $\lim_{k \to \infty} u(t_k^*, \cdot)$ exists as an element of $C^2(\bar{\Omega})$. By definition, the *limit set* Λ^+ is the set of all functions in $C^2(\bar{\Omega})$ that can be obtained as limits in this way. Characterization of Λ^+ is the object of the following investigation.

3. Statement of the main results. All the concepts needed for our principal theorem are now at hand.

THEOREM 1 (Global existence). *Assume that the assumptions* (R) *hold and that there exists a Lyapunov function* $V: D \to R$ *of class* C^2 *such that* $\dot{V}(z) \leqq 0$ *and such that* $V_{zz} \geqq \beta I_n$, *where* $\beta \geqq 0$ *is constant. Then* (6) *has exactly one solution* u *for each initial value* $\bar{u}$. *The solution exists for* $t \geqq 0$, *it belongs to* $C^*(0, \infty)$, *and it remains in a compact subset of* D. *The function* $u(t, \cdot)$ *is continuous and the orbit* $\{u(t, \cdot): 0 \leqq t < \infty\}$ *is bounded in the* $C^{2+\gamma}$ *norm.*

THEOREM 2 (Asymptotic behavior). *Under the assumptions of Theorem* 1, *with* $\beta > 0$, *the limit set* Λ^+ *of* u *is a nonempty compact subset of* $C^2(\bar{\Omega})$ *and it contains constant functions only. Considered as a subset of* R^n, *the limit set* Λ^+ *is an invariant set for the ordinary differential equation* $u' = f(u)$, *and* $\dot{V}$ *vanishes on* Λ^+.

The above results imply that

$$\operatorname{dist}(u(t, \cdot), \Lambda^+) \to 0 \quad \text{in } C^2(\bar{\Omega}) \quad \text{as } t \to \infty.$$

In particular, given any sequence $\{t_k\}$ with $t_k \to \infty$ there exists a subsequence $\{t_k^*\}$ and a constant $c \in R^n$ such that $\|u(t_k^*, \cdot) - c\|_2 \to 0$ as $k \to \infty$.

Before turning to the proof we indicate a possible source of misunderstanding. Since each element of Λ^+ is a constant c, and all the derivatives of c are of course 0, it might be thought that the partial differential equation implies $f(c) = 0$. (The uniformity of convergence does in fact give $Lu \to Lc = 0$ as $t = t_k \to \infty$.) Although existence of a "stationary point" c with $f(c) = 0$ was not postulated initially, it was pointed out by Professor Robert Greene of the University of California, Los Angeles that this is implied by the hypothesis of Theorem 2. Namely, $f(c) = 0$ at the point where V assumes its minimum; we omit the easy proof.

Nevertheless, the above conclusion "$c \in \Lambda^+ \Rightarrow f(c) = 0$" is false. Taking $n = 2$ let $u = (p, q)$ and consider the system

$$p_t = Lp + q, \qquad q_t = Lq - p, \qquad u(0, x) = (0, 1).$$

This system has the solution $u(t, x) = (\sin t, \cos t)$ and the hypothesis of Theorem 2 holds with $V(z) = |z|^2$. But no point $c \in \Lambda^+$ satisfies $f(c) = 0$.

Later we shall consider a similar problem for the scalar function $U = V(u)$ and here the elements $W \in \Lambda^+$ for U will satisfy $W_t = 0$ as well as $LW = 0$. The difference in the two cases is that in the scalar case we shall have $U(t, \cdot) \to c$ as $t \to \infty$ without restriction, while $u(t, \cdot) \to c$ holds only on the sequence $\{t_k\}$.

4. Existence of a limit. One of the most important lemmas needed for our theorem states that the solutions of $\phi_t \leqq L\phi$, $\phi_\nu \leqq 0$ have a constant limit, in general, as $t \to \infty$. In [14] the result is deduced first for the corresponding equation $\phi_t = L\phi$, $\phi_\nu = 0$ by consideration of the integrals

$$h(t) = \int_\Omega \phi(t, x)\, dx, \qquad k(t) = \int_\Omega \phi^2(t, x)\, dx.$$

Another method is briefly outlined as follows: Let (λ_n, ψ_n) be the sequence of characteristic values and functions for the problem

$$L\psi + \lambda\psi = 0, \qquad \psi_\nu = 0.$$

Then $\lambda_1 = 0$, $\psi_1 = 1$, and $\lambda_n > 0$ for $n > 1$. Also, by the spectral theorem,

$$\phi(0, x) = \sum_{n=1}^{\infty} b_n \psi_n(x) \Rightarrow \phi(t, x) = \sum_{n=1}^{\infty} b_n\, e^{-\lambda_n t} \psi_n(x),$$

where the coefficients b_n are bounded. Hence $\phi(t, x) \to b_1$ as $t \to \infty$, and indeed, with exponential rapidity. A comparison argument [14] now gives the result for the inequality, which we formulate as follows.

LEMMA 1. *Let ϕ be a bounded real-valued function of class $C^*(0, \infty)$ satisfying*

$$\phi_t \leqq L\phi, \quad \phi_\nu \leqq 0, \qquad \sup |\phi_t| < \infty, \qquad \sup |\phi_x| < \infty.$$

Then $\lim_{t\to\infty} \phi(t, x) = c$, where c is constant, and the convergence is uniform with respect to x.

Since the hypothesis on ϕ_t and ϕ_x in Lemma 1 forms the primary obstacle to extension of our theorem to quasilinear systems, we shall give an alternative proof, different from that in [14], which requires weaker assumptions and makes no use of existence theory. It will be seen that the crucial requirement is a condition of uniform continuity of ϕ, and, if the assertion of uniformity of the convergence is dropped, even this is needed only in each compact subset of Ω as $t \to \infty$.

5. Proof of Lemma 1. Let $\phi = M - e^{-\psi}$ where M is a sufficiently large constant; for example, $M = 1 + \sup \phi$. Then with $\phi_j = D_j\phi$, $\phi_{ij} = D_iD_j\phi$,

$$\phi_j = e^{-\psi}\psi_j, \qquad (a_{ij}\phi_j)_i = e^{-\psi}[(a_{ij}\psi_j)_i - a_{ij}\psi_i\psi_j]$$

and hence

$$\phi_t = e^{-\psi}\psi_t \leqq L\phi = e^{-\psi}(L\psi - a_{ij}\psi_i\psi_j)$$

or, in view of the hypothesis $(a_{ij}) \geqq \alpha I$,

$$\psi_t \leqq L\psi - \alpha|\psi_x|^2.$$

Let us now set

$$h(t) = \frac{1}{|\Omega|} \int \psi(t, x)\, dx.$$

Then by the divergence theorem

$$h'(t) \leqq -\frac{\alpha}{|\Omega|} \int |\psi_x(t, x)|^2\, dx.$$

It is readily checked that e^{ψ} (as well as $e^{-\psi}$) is bounded, and hence the boundedness of ϕ_t and ϕ_x ensures that ψ_t and ψ_x are also bounded. From the regularity conditions on Ω it follows that ψ is uniformly Lipschitz continuous. That is, there exists a constant L such that

$$|\psi(t, x)-\psi(s, y)| \leqq L(|s-t|+|x-y|) \quad \text{for } x, y \in \bar{\Omega} \text{ and } s, t \geqq 0.$$

Lemma 1 is therefore a consequence of Lemma 2.

LEMMA 2. *Assume that ψ is bounded and uniformly continuous in $(0, \infty)\times\Omega$ and that $\psi_x(t, \cdot) \in L_2(\Omega)$ for $t>0$. Suppose*

$$h(t)=\frac{1}{|\Omega|}\int \psi(t, x)\, dx \quad \textit{satisfies } h'(t) \leqq -\lambda \int |\psi_x(t, x)|^2\, dx,$$

where λ is a positive constant. Then $\lim_{t\to\infty} \psi(t, x)=c$, where c is constant, and the convergence is uniform with respect to x.

Proof of Lemma 2. Let

$$v(t, B)=\sup_{x\in B} \psi(t, x)-\inf_{x\in B} \psi(t, x) \quad \text{where } B \subset \Omega.$$

We show first that $v(t, B)\to 0$ as $t\to\infty$ for any closed ball $B\subset\Omega$. Assume that B has radius r and that $\operatorname{dist}(B, \Gamma)=\alpha>0$. Suppose that for a specified value t_0 we have $v(t_0, B)>3\varepsilon$. Then there are points $\xi, \eta \in B$ such that $|\psi(t_0, \xi)-\psi(t_0, \eta)|>2\varepsilon$. Since ψ is uniformly continuous there exists $\delta \leqq \alpha$ (independent of t_0) with the property that

$$|\psi(t, x)-\psi(t, y)|>\varepsilon \quad \text{for } |t-t_0|, |x-\xi|, |y-\eta|<\delta.$$

Assume for simplicity that $\xi_1 \neq \eta_1$ and let $p=(0, p')$, where $p'=(p_2, \cdots, p_n) \in R^{n-1}$, $|p'|<\delta$. Then

$$\varepsilon<|\psi(t, \xi+p)-\psi(t, \eta+p)| \leqq \int_{\xi_1}^{\eta_1} |\psi_x(t, s, p_2, \cdots, p_n)|\, ds.$$

Integrating over the ball B': $|p'|<\delta$ in R^{n-1}, we obtain

$$|B'|\varepsilon<\int_{B'}\int_{\xi_1}^{\eta_1} |\psi_x(t, x)|\, dx \leqq \left(\int_{\Omega} |\psi_x(t, x)|^2\, dx \cdot 2r|B'|\right)^{1/2},$$

where $|B'|=\omega_{n-1}\delta^{n-1}$ is the $(n-1)$-dimensional volume of B'. This inequality shows that

$$\int |\psi_x(t, x)|^2\, dx \geqq \frac{|B'|}{2r}\varepsilon^2 =: \beta \quad \text{for } (t-t_0)<\delta.$$

Hence $h' \leqq -\lambda\beta$, i.e., $h(t_0+\delta)-h(t_0-\delta)<-2\delta\lambda\beta$. If there existed a sequence (t_k), $t_k\to\infty$, such that $v(t_k, B)>3\varepsilon$, we would have $\lim_{t\to\infty} h(t)=-\infty$. Since h is bounded, this cannot happen, i.e., $v(t, B)<3\varepsilon$ for large t.

Now we show that $v(t, \Omega)\to 0$. Let $\varepsilon>0$ be given and let $\delta>0$ be such that $|\psi(t, x)-\psi(t, y)|<\varepsilon$ for $|x-y|<\delta$ and all t. Let B be a finite subset of Ω with the property that the δ-neighborhood of B covers Ω. By connecting the points of B with polygonal lines, we obtain a connected subset C of Ω with positive distance from Γ, say, $\operatorname{dist}(C, \Gamma)=2\alpha>0$. We cover C by a finite number of balls B_i with radius α and midpoint on C with the property that for each B_i there is a neighboring B_j such that $B_i \cap B_j \neq \varnothing$. Because $C_1 \cap C_2 \neq \varnothing$ implies $v(t, C_1 \cup C_2) \leqq v(t, C_1)+v(t, C_2)$, we get $v(t, C)\to 0$ as $t\to\infty$. Since each point of Ω is in the δ-neighborhood of C,

$$v(t, \Omega)<v(t, C)+2\varepsilon,$$

which shows that $v(t, \Omega)\to 0$ as $t\to\infty$.

Since h is decreasing and bounded, there exists

$$c = \lim_{t \to 0} h(t).$$

On the other hand, since $h(t)$ is a mean value, there exists $x_t \in \Omega$ such that $\psi(t, x_t) = h(t)$. The equation

$$|\psi(t, x) - h(t)| = |\psi(t, x) - \psi(t, x_t)| \leqq v(t)$$

now shows that $\psi(t, x) \to c$, and that the convergence is uniform with respect to x. This completes the proof of Lemma 2.

6. A matrix inequality. The trace of a square matrix A, written tr A, is the sum of its diagonal elements. According to a well-known lemma of Schur and Fejér, the following holds: Let A and D be square matrices of the same size with $A \geqq 0$, $D \geqq 0$, $D^T = D$. Then tr $AD \geqq 0$, with strict inequality if $A > 0$ and $D \neq 0$. Since this result is usually quoted for symmetric matrices only, let us briefly recall the proof. From $D \geqq 0$ and $D^T = D$ follows a representation $D = X^T X$, so that $d_{ij} = x_{ki} x_{kj}$. Thus,

$$\text{tr } AD = a_{ij} d_{ij} = a_{ij} x_{ki} x_{kj}.$$

For each fixed k the quadratic form on the right is $\geqq 0$, with strict inequality for at least one k if $A > 0$ and $X \neq 0$. This completes the proof.

We shall establish the following lemma.

LEMMA 3. *Let $A = (a_{ij})$, $B = (b_{ij})$ and $C = (c_{ij})$ be matrices of size m by m, n by n and n by m, respectively. Suppose further that $B = B^T$ and that $A \geqq \alpha I_m$, $B \geqq \beta I_n$ with $\alpha \geqq 0$ and $\beta \geqq 0$. Then*

$$\text{tr } AC^T BC = a_{ij} b_{kl} c_{ik} c_{jl} \geqq \alpha\beta |C|^2$$

where $|C|^2 = \sum_{i,j} (c_{ij})^2$.

For proof, let $D = C^T BC$. Then $x^T Dx = z^T Bz$ where $z = Cx$. This shows $D \geqq 0$ and the inequality tr $AC^T BC \geqq 0$ follows from the lemma of Schur and Fejér as quoted above. Now write $A = \bar{A} + \alpha I_m$, $B = \bar{B} + \beta I_n$ with $\bar{A} \geqq 0$ and $\bar{B} \geqq 0$ and consider the identity

$$(\bar{A} + \alpha I_m) C^T (\bar{B} + \beta I_n) C = \bar{A} C^T \bar{B} C + \alpha C^T \bar{B} C + \beta \bar{A} C^T C + \alpha\beta C^T C.$$

It follows from the earlier result that the traces of the first three matrices on the right are nonnegative, and hence

$$\text{tr } AC^T BC \geqq \alpha\beta \text{ tr } C^T C = \alpha\beta |C|^2.$$

7. The Lyapunov function. Let $V: D \to R$ be of class C^2 and let u be a solution of $u_t = Lu + f(u)$. The notation $V_k = \partial V(z)/\partial z_k$, $u_i^k = \partial u^k / \partial x_i$ are used, and similarly for higher derivatives. The gradient and Hessian of V are denoted by V_z and V_{zz}, respectively, so that $V_{zz} = \text{matrix } (V_{kl})$. As before, i and j run from 1 to n while k and l run from 1 to m. We are going to derive a differential inequality for the scalar function $U(t, x) = V(u(t, x))$ or, briefly, $U = V(u)$.

From

$$U_t = V_k u_t^k = V_k (a_{ij} u_i^k)_j + V_k f^k(u),$$

$$L(U) = (a_{ij} V_k u_i^k)_j = V_k (a_{ij} u_i^k)_j + a_{ij} V_{kl} u_i^k u_j^l,$$

it follows that

$$U_t = L(U) + \dot{V} - a_{ij} V_{kl} u_i^k u_j^l$$

where $\dot{V} = V_x \cdot f$ along the trajectory u. Applying Lemma 3 with $A = (a_{ij})$, $B = V_{xx}$ and $C = (u_i^k)$, we get the following lemma.

LEMMA 4. *Let $V: D \to R$ be a function of class C^2 satisfying*

$$V_{zz} \geqq \beta I, \qquad \dot{V} = V_z \cdot f(u) \leqq 0$$

where β is a positive constant and where u is a solution of $u_t = Lu + f(u)$. Then the scalar function $U = V(u)$ satisfies

$$U_t \leqq LU - \alpha\beta |u_x|^2.$$

It should be observed that the proof, based on Lemma 3, uses the fact that V_{zz} is symmetric, but does not require symmetry of the matrix A. The latter condition would be a serious restriction on the operator L.

8. Monotonicity and comparison. The following two lemmas are well known [14], [19] but are repeated here for logical completeness. They hold under much weaker assumptions regarding L, Ω and u than those necessary for existence theory. Rather than fully exploiting this fact, we point to it by introducing the space $C_0^*(J)$ of functions continuous in $\bar{J} \times \bar{\Omega}$ with derivatives u_t, u_x and u_{xx} continuous in $J \times \Omega$. As always, J is an interval $(0, T]$ or $(0, \infty)$.

LEMMA 5. *If the real-valued functions ϕ, $\psi \in C_0^*(J)$ satisfy*

$$\phi_t \leqq L\phi, \quad \psi_t \geqq L\psi, \quad \frac{\partial \phi}{\partial \nu} \leqq \frac{\partial \psi}{\partial \nu}, \quad \phi(0, x) \leqq \psi(0, x)$$

then $\phi \leqq \psi$ in $\bar{J} \times \bar{\Omega}$.

LEMMA 6. *If two solutions v, $u \in C_0^*(J, R^n)$ of $u_t = Lu + f(u)$ satisfy*

$$|u(0, x) - v(0, x)| \leqq \rho, \qquad \frac{\partial u}{\partial \nu} = \frac{\partial v}{\partial \nu}$$

and if for these solutions $|f(u) - f(v)| \leqq M|u - v|$, where M and ρ are constant, then $|u(t, x) - v(t, x)| \leqq \rho e^{Mt}$ in $\bar{J} \times \bar{\Omega}$.

9. Proof of Theorem 1. Let $\max V(\bar{u}(x)) = \eta$, where $\bar{u}$ is the initial value of u. Since Lemma 4 gives $U_t \leqq LU$, and since $\partial U / \partial \nu = 0$, we can apply Lemma 5 with $\phi = U$ and $\psi = \eta$. It follows that, as long as the solution exists, $U(t, x) \leqq \eta$ and hence $u(t, x) \in D_\eta$, where $D_\eta = \{z \in R^n : V(z) \leqq \eta\}$ is a compact subset of D (cf. the definition of Lyapunov function given in § 1). Theorem 1 now follows from Theorem 0.

10. Proof of Theorem 2. It is clear from the global estimate given in Theorem 1 that the set Λ^+ is nonempty and compact. Since $U_t \leqq LU$ by Lemma 4, and since the differential equation together with Theorem 0 shows that u_t and hence U_t are bounded, we can apply Lemma 1 to get

$$\lim_{t \to \infty} U(t, \cdot) = c, \qquad c \text{ constant.} \tag{7}$$

Let $\bar{w}$ be an arbitrary element of Λ^+, say $\lim u(t_k, \cdot) = \bar{w}$ as $k \to \infty$. Clearly $\bar{w}$ satisfies (R4). We denote by $w(t, x)$ the solution of (6) with initial value $\bar{w}$ and we set $W(t, x) = V(w(t, x))$, that is, $W = V(w)$. The proof of Theorem 1 gives $\bar{w}(x) \in D_\eta$, hence $w(t, x) \in D_\eta$ by the same proof, and we may assume, therefore, that $|f(u) - f(w)| \leqq M|u - w|$ for some constant M. From Lemma 6 we obtain the estimate

$$|u(t_k + t, x) - w(t, x)| \leqq \|u(t_k, \cdot) - \bar{w}\|_0 \, e^{Mt} \tag{8}$$

which implies $u(t_k+t, x)\to w(t, x)$ and $U(t_k+t, x)\to W(t, x)$ uniformly in x as $k\to\infty$. But the latter limit is constant by (7); hence $W(t, x)=c$ for all t and x. Now, Lemma 4 gives

$$W_t \leqq LW-\alpha\beta|w_x|^2$$

and, on the other hand, $W_t=LW=0$. Hence $w_x=0$. This shows that $w(t, x)$ and in particular $w(x)=w(0, x)$ are independent of x. In other words, $w=w(t)$ is a solution to the ordinary differential equation $w_t=f(w)$. It follows from (8) that $w(t)\in\Lambda^+$ for any positive t; in other words, Λ^+ is positively invariant. That Λ^+ is also negatively invariant, and that $\dot{V}(z)$ vanishes on Λ^+, are proved as in [14].

11. Remarks on the condition for V_{zz}. If the condition $V_{zz}>0$ is replaced by $V_{zz}\geqq 0$ Theorem 2 no longer follows even when $m=n=1$. To see this, let $f(u)$ be a smooth function such that $f(u)=u$ for $|u|\leqq 1$ and $f(u)=0$ for $|u|\geqq 2$. Also let $V(z)=0$ for $|z|\leqq 2$, $V(z)=(2-|z|)^4$ for $|z|\geqq 2$. Then the boundary value problem suggested for $n=m=1$ by

$$u_t=u_{xx}+f(u), \qquad (t, x)\in(0, \infty)\times(0, \pi)$$

has a solution $u(t, x)=\cos x$, which violates the conclusion of Theorem 2.

In many developments of the Lyapunov theory $V(z)$ is required to satisfy the additional condition

$$V(0)=0, \qquad V(z)>0 \quad \text{for } z\neq 0. \tag{9}$$

The first of these can be attained by adding a constant to V but the second is more restrictive. The condition is imposed to ensure stability of the zero solution, a consideration which has nothing to do with the problems addressed in this paper. (If we impose the additional condition $\dot{V}(z)<0$ for $z\neq 0$, as well as (9), then every solution with initial values satisfying $\bar{u}(x)\in D$ for $x\in\bar{\Omega}$ converges to the null solution in the $C^2(\Omega)$ norm, as $t\to\infty$, and the latter solution is stable. But this case excludes most of the more interesting applications of the La Salle stability theorem and it can be obtained in a much simpler way.)

If V is normalized so $\min V(z)=V(0)=0$, the hypothesis $V_{zz}>0$ ensures (9) automatically. The question arises whether (9) is sufficient for Theorem 2, without any condition on V_{zz}. We shall show that this is not the case. Considering the same equation as above, let f be a smooth function satisfying

$$f(u)=0, \quad u-4, \quad 0$$

on the intervals $u\leqq 2$, $3\leqq u\leqq 5$, $u\geqq 6$, respectively. Let $V(z)$ be a smooth function satisfying $V(z)>0$ for $z=0$ and $V(z)=z^2$, 2, z^2 on the intervals $z\leqq 1$, $2\leqq z\leqq 6$, $z\geqq 7$, respectively. Then all assumptions hold, except $V_{zz}>0$. Nevertheless the boundary value problem with initial value $\bar{u}(x)=4+\cos x$ has the solution $u=\bar{u}$ and Λ^+ contains only the function $\bar{u}$, which is certainly not constant. If $m=n=1$, a brief investigation suggests that $V_{zz}\geqq 0$ together with (9) is sufficient. Whether this holds in general is left as an open problem.

If $V(z)=f_1(z^1)+f_2(z^2)+\cdots+f_n(z^n)$ the condition $f_k''>0$ for $k=1, 2, \cdots, n$ ensures $V_{zz}>0$. Using this fact, one can show that Theorem 2 contains the main results of [14].

12. The quasilinear case. In this section we will indicate how to extend the above results to the quasilinear system ($k=1, 2, \cdots, n$)

$$u_t^k=\sum_{i,j=1}^{m}\frac{d}{dx_i}\left(a_{ij}(x, u)\frac{\partial u^k}{\partial x_j}\right)+f^k(u) \quad \text{in } J\times\Omega \tag{10}$$

subject to the initial and boundary condition (4). Instead of assumptions (R) we now require:

(R1′) Ω is a bounded domain in R^m with orientable boundary Γ of class C^{3-}.

(R2′) $D \subset R^n$ is open and $f: D \to R^n$ is locally of class C^{2-}.

(R3′) The matrix $A(x, u) = (a_{ij}(x, u))$ is of class $C^{3-}(\bar{\Omega} \times D)$ and $A \geqq \alpha I_m$ with $\alpha > 0$ constant.

(R4′) as (R4).

(R5′) For $x \in \Gamma$ the functions $a_{ij}(x, u)$ are independent of u.

Here, $C^{3-}(\bar{\Omega} \times D)$ denotes the space of all functions $\phi: \bar{\Omega} \times D \to R$ with Lipschitz continuous first and second partial derivatives; the symbols C^{3-} and C^{2-} are to be interpreted in a similar fashion.

Remark. It is sufficient for our purposes that (R3′) holds locally with respect to u.

Under the assumptions (R′) assertions analogous to those in Theorems 0, 1 and 2 can be proved for system (10). Hence, in particular, the asymptotic behavior for (10) is completely determined by the asymptotic behavior of the solutions to the system (1) of ordinary differential equations. Since, as has been noted above, the proofs of Theorems 1 and 2 can be easily adapted to the quasilinear setting, the main difficulty in establishing this result lies in the extension of Theorem 0 to (10).

We do not want to go into details about the proof at this place. Let us only say that local existence of a classical solution u for the boundary-value problem (10), (4) follows from [5, § 10]. Next, using reasoning similar to that employed in [16, § 4], an a priori bound (uniformly in t) for the functions $u(t, \cdot)$ in $C^{1+\alpha}(\bar{\Omega}, R^n)$ is established (any $0 < \alpha < 1$). This is sufficient to guarantee global existence of u. The statement about global boundedness of $\|u(t, \cdot)\|_{2+\gamma}$ then follows from [16, Thm. 8]. We have thus proved the following.

THEOREM 3. *Assume that the assumptions* (R′) *hold and that there exists a Lyapunov function* $V: D \to R$ *of class* C^2 *such that* $\dot{V}(z) \leqq 0$ *in* D *and such that* $V_{zz} \geqq \beta I_n$ *where* $\beta > 0$ *is constant. Then* (10), (4) *has exactly one solution* u *for each initial value* $\bar{u}$. *The solution exists for* $t \geqq 0$, *it belongs to* $C^*([0, \infty))$ *and it remains in a compact subset of* D. *The orbit* $\{u(t, \cdot): t \geqq 0\}$ *is bounded in the* $C^{2+\gamma}$ *norm.*

The limit set Λ^+ *of* u *is a nonempty compact subset of* $C^2(\bar{\Omega})$ *and it contains constant functions only. Considered as a subset of* R^n, Λ^+ *is an invariant set for the ordinary differential equation* $u' = f(u)$ *and* $\dot{V}$ *vanishes on* Λ^+.

13. Lyapunov functionals. We briefly describe an alternative approach. Assume that u is a bounded regular solution of (2), (4) with values in D, and that $V: D \to R$ is a Lyapunov function for the associated system (1) of ordinary differential equations with $V_{zz} \geqq \beta I_n$, where $\beta > 0$ is constant. Consider the Lyapunov functional

$$U(t) = \int_\Omega V(u(t, x))\, dx \quad \text{for } t \geqq 0.$$

We get

$$\begin{aligned} \dot{U}(t) = \frac{d}{dt} U(t) &= \int_\Omega V_k u_t^k \\ &= \int_\Omega V_k (a_{ij} u_j^k) i + V_k f^k \\ &= \int_\Omega V_k f^k - a_{ij} V_{kl} u_j^k u_i^l \\ &\leqq \int_\Omega V_k f^k - \alpha\beta |u_x|^2. \end{aligned}$$

Since the orbit $\{u(t,\cdot): t \geqq 0\}$ is relatively compact in $C^2(\bar{\Omega}, R^n)$ by [15], La Salle's principle implies that Λ^+ consists of constant functions only. Moreover, Λ^+ is an invariant set for (1) and $\dot{V}=0$ on Λ^+. This proves Theorem 2.

The same reasoning also applies to the system

$$(2^*) \qquad u_t^k = L^k u^k + f^k(u) \quad \text{in } J\times\Omega \quad (k=1,2,\cdots,n),$$

where the elliptic part

$$L^k\theta = (a_{ij}^k(x)\theta_j)_i$$

now may vary with k, provided we have $a_{ij}^k = c^k(x)a_{ij}(x)$ with functions $c^k(x) \in C^{1+\gamma}(\bar{\Omega})$ that are strictly positive in $\bar{\Omega}$. Moreover we suppose that

(α) (a_{ij}) is symmetric and $(c_k V_{kl})$ is positive definite for any $x \in \bar{\Omega}$, $u \in D$;

or that

(β) the matrix V_{zz} is of diagonal form.

This gives the following theorem.

THEOREM 2*. *Let u be a bounded regular solution of* (2), (4) *under the above hypotheses. Then the limit set Λ^+ of u is a nonempty compact subset of $C^2(\bar{\Omega})$ and it contains constant functions only. Moreover, Λ^+ is an invariant set for* (1) *and $\dot{V}$ vanishes on Λ^+.*

Theorem 2* generalizes a result of Alikakos [1, § 4], where the case $(a_{ij}) = I_m$, c_k constant is considered. The method of proof (integration with respect to some variables and derivation of a differential inequality for the resulting function) apparently was first used by Carleman [6] in his paper on the Denjoy conjecture.

14. Remarks on boundedness. It should be noted that in the preceding section boundedness of the solution u is used as a hypothesis. As shown above, for systems of type (2) an a priori bound for u can be derived directly from a differential inequality for the Lyapunov function. For system (2*) it seems that such a procedure is not possible. We mention instead the following two methods.

(i) *Comparison arguments.* Let $\alpha(t)$, $\beta(t):[0,\infty)\to R^n$ be a pair of upper and lower solutions for (2*), i.e., assume that $(k=1,2,\cdots,n;\ t>0)$

$$\alpha_t^k(t) \leqq f^k(z) \quad \text{for all } \alpha(t) \leqq z \leqq \beta(t), \quad z^k = \alpha^k(t),$$

$$\beta_t^k(t) \geqq f^k(z) \quad \text{for all } \alpha(t) \leqq z \leqq \beta(t), \quad z^k = \beta^k(t),$$

and that

$$\alpha(0) \leqq u(0,x) \leqq \beta(0) \quad \text{in } \bar{\Omega}.$$

Then $\alpha(t) \leqq u(t,x) \leqq \beta(t)$ for all $t \geqq 0$, $x \in \bar{\Omega}$. This is a special case of a general estimation theorem for parabolic systems (see [19, § 32]). In case α and β are constant it is sometimes called the method of "invariant rectangles."

(ii) *Functional analytic methods* (bootstrapping and feedback arguments). For these we refer the reader to the book by Rothe [18].

15. Historical remarks. As mentioned in the Introduction, a number of special cases of (2) have been treated in the literature. Our results contain those of [10], [12] and [14], and they generalize those of [20] from one to several space variables. The results of [17] are contained insofar as they deal with Neumann boundary conditions. Our results do not contain those of [17] for other boundary conditions, nor those of [7], [8], [9] or [13]. Reference [7] pertains to an unbounded x-domain, [8] and [9] involve nonlinearities depending on both x and u, and [13] has a functional acting on $u(t,\cdot)$. These references suggest possible directions in which our work might be

generalized. It would also be desirable to extend the results to equations with a lagging time variable.

REFERENCES

[1] N. D. ALIKAKOS, *An application of the invariance principle to reaction-diffusion equations*, J. Differential Equations, 33 (1979), pp. 201-225.

[2] ———, *Remarks on invariance in reaction-diffusion equations*, J. Nonlinear Analysis, 5 (1981), pp. 593-614.

[3] ———, *A Lyapunov functional for a class of reaction-diffusion systems*, Conference Southern Illinois Univ., Carbondale, IL, 1978; Dekker Lecture Notes 58, Marcel Dekker, 1980.

[4] H. AMANN, *Invariant sets and existence theorems for semilinear parabolic and elliptic systems*, J. Math. Anal. Appl., 65 (1978), pp. 432-467.

[5] ———, *Quasilinear evolution equations and parabolic systems*, Trans. Amer. Math. Soc., 293 (1986), pp. 191-228.

[6] T. CARLEMAN, *Sur une inégalité différentielle dans la théorie des fonctions analytiques*, Compt. Rendus, 196 (1933), pp. 995-997.

[7] N. CHAFFEE, *A stability analysis for a semilinear parabolic partial differential equation*, J. Differential Equations, 15 (1974), pp. 522-540.

[8] ———, *Asymptotic behavior for solutions of a one-dimensional parabolic equation with homogeneous Neumann boundary conditions*, J. Differential Equations, 18 (1975), pp. 111-134.

[9] W. H. FLEMING, *A selection-migration model in population genetics*, J. Math. Biol., 2 (1975), pp. 219-233.

[10] K. P. HADELER, *Diffusion in Fisher's population model*, Rocky Mountain J. Math., 11 (1981), pp. 39-45.

[11] J. P. LASALLE, *Stability theory for ordinary differential equations*, J. Differential Equations, 4 (1968), pp. 57-65.

[12] A. LEUNG, *Limiting behavior for a prey-predator model with diffusion and crowding effects*, J. Math. Biol., 6 (1978), pp. 87-93.

[13] P. MOTTONI, E. ORLANDI AND A. TESEI, *Asymptotic behavior for a system describing epidemics with migration and spatial spread of infection*, Nonlinear Analysis TMA, 3 (1979), pp. 663-675.

[14] R. REDHEFFER AND W. WALTER, *On parabolic systems of the Volterra prey-predator type*, J. Nonlinear Analysis, 7 (1983), pp. 333-347.

[15] R. REDLINGER, *Über die C^2-Kompaktheit der Bahn von Lösungen semilinearer parabolischer Systeme*, Proc. Roy. Soc. Edinburgh, Sect. A, 93 (1982), pp. 99-103.

[16] ———, *Compactness results for time-dependent parabolic systems*, J. Differential Equations, 64 (1986), pp. 133-153.

[17] F. ROTHE, *Convergence to the equilibrium state in the Volterra-Lotka diffusion equations*, J. Math. Biol., 3 (1976), pp. 319-324.

[18] ———, *Global Solutions of Reaction-Diffusion Systems*, Lecture Notes in Mathematics 1072, Springer-Verlag, Berlin, 1984.

[19] W. WALTER, *Differential and integral inequalities*, Ergeb. der Math. u. ihrer Grenzgebiete Bd. 55, Springer-Verlag, Berlin, 1970.

[20] G. F. WEBB, *A reaction-diffusion model for a deterministic diffusive epidemic*, J. Math. Anal. Appl., 84 (1981), pp. 150-161.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

010

BOUNDED SOLUTIONS OF VOLTERRA EQUATIONS*

JAN PRÜSS†

Abstract. The solvability behavior on the real line of linear integrodifferential equations in a general Banach space is considered and several applications to integral partial differential equations are given.

Key words. Volterra equations, Laplace transform, Fourier transform, Fourier-Carleman transform, admissibility, almost periodic solution, resolvents

AMS(MOS) subject classifications. 45N05, 45K05, 42A38

1. Introduction. It is well known that quite a few problems of applied mathematics lead to abstract equations of the following type:

$$u'(t) = Au(t) + \int_0^\infty dB(\tau)u(t-\tau) + f(t), \tag{1}$$

for example, problems in thermodynamics or elasticity theory for materials with memory, and in population dynamics, to mention a few. Here $u(t)$ denotes the state of the system at time t, A a closed linear operator in the state space, a Banach space X, with dense domain $D(A)$, $\{B(t)\}_{t\geqq 0}$ a family of closed linear operators in X with $D(B(t)) \supset D(A)$ for all $t \geqq 0$ such that $B \in BV(\mathbb{R}_+, \mathbb{B}(Y, X))$ (the space of $\mathbb{B}(Y, X)$-valued functions of bounded variation over $\mathbb{R}_+ = [0, \infty)$), where $Y = D(A)$ is normed by the graph norm of A and $\mathbb{B}(Y, X) = \{T: Y \to X : T \text{ linear and bounded}\}$, and f an X-valued function; w.l.o.g. we also assume that $B(\cdot)$ is left-continuous in $\mathbb{B}(Y, X)$ and satisfies $B(0) = B(0+) = 0$.

In recent years equations of type (1) have been the object of intensive study, mainly the local problem for (1) has been investigated, i.e., the history value problem. This means, given the history function $u_-(t)$ for $t \leqq 0$ and f, find the solution u such that $u(t) = u_-(t)$ for $t \leqq 0$ and (1) holds for $t \geqq 0$. Assuming well-behaved history values $u_-(t)$, this problem reduces to the initial value problem for

$$u'(t) = Au(t) + \int_0^t dB(\tau)u(t-\tau) + g(t), \qquad t \in \mathbb{R}_+, \tag{2}$$

where

$$g(t) = f(t) + \int_t^\infty dB(\tau)u_-(t-\tau).$$

The theory of (2) centers around the concepts "wellposedness" and "resolvent operator" and is well understood by now. It turns out that (2) is well posed if and only if there is a resolvent operator $S(t)$ for (2), and in this case the solution of (2) with initial value u_0 is represented by the variation of parameters formula

$$u(t) = S(t)u_0 + \int_0^t S(t-\tau)g(\tau)\, d\tau. \tag{3}$$

Furthermore, the relation

$$\hat{S}(\lambda) = (\lambda - A - \widehat{dB}(\lambda))^{-1} \tag{4}$$

* Received by the editors September 2, 1986; accepted for publication November 6, 1986.
† Fb 17, UHGS Paderborn, Warburger Str. 100, 4790 Paderborn, Federal Republic of Germany.

for the Laplace-transform of S yields a Hille-Yosida type result for resolvents analogous to linear differential equations (cf., e.g., Grimmer and Prüss [7] or Prüss [13]).

However, there are other questions concerning (1) that call for a global theory of (1), e.g., the existence of periodic or almost periodic solutions in the case f has the corresponding property. It is the purpose of this note to present some results in this direction.

2. Admissibility. Let $f:\mathbb{R}\to X$ be continuous, $f\in C(X)$ for short. First we have to state what we mean by a solution of (1).

DEFINITION 1. $u\in C(X)$ is called a *solution* of (1) if $u\in C(Y)\cap C^1(X)$, $\sup_{t\leqq 0}|u(t)|_Y<\infty$ and (1) holds on $\mathbb{R}$.

Note that (1) is translation invariant; i.e., if u is a solution of (1), then $(T_\tau u)(t)=u(t+\tau)$ is also a solution of (1) with f replaced by $T_\tau f$. This property is not shared by (2). The analogue of "wellposedness" for (1) is the concept of an admissible subspace of $C_u(X)=\{f\in C(X): f$ bounded and uniformly continuous$\}$ which we have to introduce next. Since (1) is translation invariant, we only consider translation invariant subspaces $W\subset C_u(X)$.

DEFINITION 2. Let $W\subset C_u(X)$ be a closed translation invariant subspace. W is called *admissible* for (1), if for each $f\in W_0=W\cap C_u^1(X)$ there is a unique solution $u\in W_0\cap C_u(Y)$, and $(f_n)\subset W_0$, $f_n\to 0$ in W imply $u_n\to 0$ in W.

In case W is admissible for (1), the solution operator G is defined on the dense set W_0 according to the following:

$$(Gf)(t)=u(t),\qquad t\in\mathbb{R},\quad f\in W_0 \tag{5}$$

and can be extended to all of W, since G is bounded, i.e., $G\in\mathbb{B}(W)$. Due to the translation invariance of (1) and W, G enjoys the following properties.

(i) $T_\tau G=GT_\tau$ for all $\tau\in\mathbb{R}$;

(ii) $DG=GD$, where $D=d/dt$;

(iii) $G(\varphi * f)=\varphi * Gf$ for all $f\in W$, $\varphi\in L^1$.

Property (i) shows already that for $f\in W$ periodic or almost periodic the solution $u=Gf$ is periodic or almost periodic, also. Therefore the main problem is to obtain characterizations of admissibility of subspaces W which can be checked more easily. We are not able to do this for all subspaces W but for a large class which we introduce next.

DEFINITION 3. Let $f\in C_b(X)=\{f\in C(X): f$ bounded$\}$. The *Fourier-Carleman* transform of f is defined by

$$\hat f(\lambda)=\begin{cases}\displaystyle\int_0^\infty e^{-\lambda t}f(t)\,dt, & \operatorname{Re}\lambda>0,\\[2mm] \displaystyle-\int_{-\infty}^0 e^{-\lambda t}f(t)\,dt, & \operatorname{Re}\lambda<0;\end{cases} \tag{6}$$

$\hat f(\lambda)$ is holomorphic in $\mathbb{C}\setminus i\mathbb{R}$. Let $\rho(f)$ denote the set of all $\rho\in\mathbb{R}$ such that $\hat f(\lambda)$ admits analytic continuation to some ball $B_\varepsilon(i\rho)$; then $\sigma(f)=\mathbb{R}\setminus\rho(f)$ is called the *spectrum* of f.

For $\Lambda\subset\mathbb{R}$ closed we let

$$\Lambda(X)=\{f\in C_u(X):\sigma(f)\subset\Lambda\}; \tag{7}$$

it can be shown that $\Lambda(X)$ is a closed translation invariant subspace of $C_u(X)$ (cf. Katznelson [11]). This class of subspaces is, on the one hand, large enough for our purposes and, on the other hand, amenable to analysis, in particular to transform

theory, and therefore presents an efficient means to describe the solvability behavior of (1). Two special cases should be mentioned.

(i) $f \in C_b(X)$ is τ-periodic. Then

$$\hat{f}(\lambda) = (1 - e^{\lambda\tau})^{-1} \int_0^\tau e^{-\lambda t} f(t)\, dt \quad \text{for all } \operatorname{Re}\lambda \neq 0;$$

hence,

$$\sigma(f) = \left\{ \frac{2\pi n}{\tau} : n \in \mathbb{Z}, f_n \neq 0 \right\} \quad \text{where } f_n = \tau^{-1} \int_0^\tau e^{2\pi\, \mathrm{int}/\tau} f(t)\, dt$$

denotes the nth *Fourier-coefficient* of f.

(ii) $f \in C_b(X) \cap L^1(X)$. Then $\sigma(f) = \operatorname{supp} \tilde{f}$, where $\tilde{f}$ denotes the *Fourier-transform* of f; this generalizes to any $f \in C_b(X)$ if $\tilde{f}$ is understood in the sense of distributions: $\sigma(f) = \operatorname{supp} \tilde{D}_f$ (cf. Katznelson [11]).

We are now in a position to state the conditions necessary for the admissibility of the subspaces $\Lambda(X)$. In fact, *Fourier-transformation* of (1) formally yields the relation

$$\begin{aligned} &(i\rho - A - \widehat{dB}(i\rho))\tilde{u}(\rho) = \tilde{f}(\rho), \qquad \rho \in \mathbb{R}, \quad \text{i.e.,} \\ &\tilde{u}(\rho) = (i\rho - A - \widehat{dB}(i\rho))^{-1}\tilde{f}(\rho) = H(i\rho)\tilde{f}(\rho). \end{aligned} \tag{8}$$

Let

$$\Lambda_0 = \{\rho \in \mathbb{R} : i\rho - A - \widehat{dB}(i\rho) \text{ not invertible}\} \tag{9}$$

denote the (real) *spectrum* of (1); then we have

PROPOSITION 1. *Let $\Lambda(X)$ be admissible for* (1). *Then*

(i) $\Lambda \cap \Lambda_0 = \varnothing$ *and there is* $M \geqq 1$ *such that*

$$|H(i\rho)| \leqq M \quad \textit{and} \quad |AH(i\rho)| \leqq M(1 + |\rho|) \quad \textit{for all } \rho \in \Lambda;$$

(ii) $\sigma(Gf) = \sigma(f)$ *for each* $f \in \Lambda(X)$;

(iii) $\sigma(u) \subset \Lambda_0$ *for each solution* $u \in C_u^1(X) \cap C_u(Y)$ *of the homogeneous equation* (1) (*i.e.*, $f = 0$).

It is obvious that the conditions necessary for admissibility of $\Lambda(X)$ presented in Proposition 1(i) are much easier to check than the definition of admissibility, but unfortunately they are not in general sufficient as a counterexample shows, even for differential equations (cf. Greiner, Voigt and Wolff [6] and Prüss [13]).

3. The main results. In this section we present several results on the solvability of (1) and on the converse of Proposition 1(i). In particular, we obtain characterizations of the admissibility of the subspaces $\Lambda(X)$ introduced above for several important subclasses of (1). One of the main tools for this purpose is the following lemma.

LEMMA 1. *Suppose $\varphi \in L^1$ satisfies $\tilde{\varphi} \in C_0^\infty$ and $\Lambda_0 \cap \operatorname{supp} \tilde{\varphi} = \varnothing$. Then there is a unique $G_\varphi \in C_b^\infty(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ such that*

$$G_\varphi' = AG_\varphi + dB * G_\varphi + \varphi = G_\varphi A + G * dB + \varphi \tag{10}$$

holds on $\mathbb{R}$.

Fourier-transformation of (10) shows that G_φ is given by

$$G_\varphi(t) = (2\pi)^{-1} \int_{-\infty}^{\infty} H(i\rho)\tilde{\varphi}(\rho)\, e^{i\rho t}\, d\rho$$

which exists and belongs to $C_b^\infty(\mathbb{B}(X, Y))$ since φ has compact support. The main task in the proof of Lemma 1 is to show that $G_\varphi \in L^1(\mathbb{B}(X, Y))$ holds; this in turn follows from the generalization of classical Paley-Wiener Theorem which we state as follows.

LEMMA 2. *Suppose $K \in L^1(\mathbb{B}(X))$ is such that $I - \tilde{K}(\rho)$ is invertible for each $\rho \in \mathbb{R}$. Then there is a unique $L \in L^1(\mathbb{B}(X))$ such that*

$$L = K + K * L = K + L * K \tag{11}$$

holds in $L^1(\mathbb{B}(X))$.

For a proof of this result we refer to Hagedorn [9], Prüss [13] or Gripenberg [8].

By proper choice of φ Lemma 1 implies our first main result.

THEOREM 1. *Let $\Lambda \subset \mathbb{R}$ be closed and such that $\Lambda \cap \Lambda_0 = \emptyset$. Then (1) is uniquely solvable in $\Lambda(X)$ for each f from a dense subspace of $\Lambda(X)$. If, in addition, Λ is compact then $\Lambda(X)$ is admissible for (1) and there is a kernel $G_\Lambda \in C_b^\infty(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ such that the solution operator admits the representation*

$$(Gf)(t) = \int_{-\infty}^{\infty} G_\Lambda(\tau) f(t-\tau)\, d\tau, \qquad t \in \mathbb{R}, \tag{12}$$

for each $f \in \Lambda(X)$.

Theorem 1 shows that, in the case where $\Lambda \cap \Lambda_0 = \emptyset$, without any further assumption on A or B the solution operator G for (1) on $\Lambda(X)$ is already densely defined, and to obtain the admissibility of $\Lambda(X)$ it remains to prove its boundedness. The second part of Theorem 1 shows that the solution operator for (1) behaves very well on $\Lambda(X)$ for compact Λ. But this should not be surprising since each $f \in C_u(X)$ with $\sigma(f)$ compact admits an extension to an entire function, namely by $f(z) = (2\pi i)^{-1} \int_\Gamma \hat{f}(\lambda)\, e^{\lambda z}\, d\lambda$, where Γ denotes a Jordan curve surrounding the set $i\sigma(f)$ in the complex plane.

The next result on admissibility does not restrict $\Lambda \subset \mathbb{R}$ (besides $\Lambda \cap \Lambda_0 = \emptyset$ of course), but the class of operators A and $B(t)$.

THEOREM 2. *Suppose A generates an analytic semigroup and let $B \in BV(\mathbb{R}_+, \mathbb{B}(Y, X))$ be decomposed according to $B = B_1 + B_2 + B_3$ where $B_1 \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{B}(Y, X))$, $B_2 \in BV(\mathbb{R}_+, \mathbb{B}(Y^\alpha, X))$ for some $\alpha < 1$, and $B_3 \in BV(\mathbb{R}_+, \mathbb{B}(Y, X))$ has sufficiently small variation. Then $\Lambda(X)$ is admissible for (1) iff $\Lambda \cap \Lambda_0 = \emptyset$. Moreover, in this case there is a kernel $G_\Lambda \in L^1(\mathbb{B}(X)) \cap L^\infty(\mathbb{B}(X))$, strongly continuous for $t \neq 0$, satisfying the jump relation $G_\Lambda(0+) - G_\Lambda(0-) = I$, such that the solution operator G for (1) is represented by (12) on $\Lambda(X)$.*

Here $Y^\alpha = D((\omega_0 - A)^\alpha)$ denotes the domain of the fractional power $(\omega_0 - A)^\alpha$ where $\omega_0 \geqq 0$ is chosen large enough; thus the assumption on B_2 means that B_2 is of "lower order." Note that the conditions for B are strong enough to imply compactness of Λ_0 as well as $\lim_{|\rho| \to \infty} |H(i\rho)| = 0$, therefore the other necessary conditions of Proposition 1(i) are automatically satisfied. Theorem 2 covers the so-called *parabolic case*; applications are given in § 6. It is an open question whether this result holds in the case where B is merely of bounded variation in $\mathbb{B}(Y, X)$; examples show that Λ_0 then need not be compact.

Our third result is concerned with A generating merely a C_0-semigroup. Then as the above-mentioned examples in [6] and [13] show, X must be a Hilbert space for Proposition 1(i) to be sufficient for admissibility. Also we have to assume more regularity as well as a certain decay of the kernel B which already implies existence of the resolvent for (2) (cf. [7]). In this case Λ_0 need not be compact as before, and here it still remains open whether the solution operator G admits a representation (12).

However, if in addition compactness of Λ_0 is assumed, then we show that such a kernel G_Λ does exist.

THEOREM 3. *Let $\Lambda \subset \mathbb{R}$ be closed, X a Hilbert space, A a generator of a C_0-semigroup in X and let $B \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}_+, \mathbb{B}(Y, X))$ be such that*

(i) $|B'(t)x| \leqq b(t)|x|_Y$ *for* $x \in D(A)$, $\int_0^\infty (1+t)b(t)\,dt < \infty$;

(ii) $B'(\cdot)x \in BV(\mathbb{R}_+, X)$ *and* $\int_0^\infty (1+t)|dB'(t)x| < \infty$ *for each* $x \in D(A)$.

Then $\Lambda(X)$ is admissible for (1) iff $\Lambda \cap \Lambda_0 = \varnothing$ and there is some $M \geqq 1$, such that $|H(i\rho)| \leqq M$ and $|AH(i\rho)| \leqq M(1+|\rho|)$ for all $\rho \in \Lambda$. In this case for each $x \in X$ there is $g_x \in L^1(X) \cap L^2(X)$ such that $\tilde{g}_x(\rho) = H(i\rho)x$ holds for all $\rho \in \Lambda$.

If, in addition, Λ_0 is compact then there is $G_\Lambda : \mathbb{R} \to \mathbb{B}(X)$ strongly continuous for $t \neq 0$, such that $G_\Lambda(\cdot)x \in L^1(X) \cap L^\infty(X)$, $G_\Lambda(0+)x - G_\Lambda(0-)x = x$ and $\tilde{G}_\Lambda(\rho)x = H(i\rho)x$ on Λ for each $x \in X$.

4. Almost periodic solutions. Recall that a function $f \in C_b(X)$ is called *almost periodic* (a.p.) if $T_{\mathbb{R}} f = \{f(\tau + \cdot) : \tau \in \mathbb{R}\} \subset C_b(X)$ is relatively compact; the space $AP(X)$ of all a.p. functions is a closed subspace of $C_b(X)$. a.p. functions are uniformly continuous on $\mathbb{R}$ and their *Bohr-transform*

$$a(\rho, f) = \lim_{N \to \infty} N^{-1} \int_0^N e^{-i\rho t} f(t)\,dt, \qquad \rho \in \mathbb{R}$$

is well defined. The *exponent set* of f

$$\exp(f) = \{\rho \in \mathbb{R} : a(\rho, f) \neq 0\}$$

is at most countable and we have $\sigma(f) = \overline{\exp(f)}$.

This can be proved by means of Bochner's approximation theorem, which states that, given a fixed countable exponent set $\{\rho_j\}_1^\infty$, there are convergence factors $\gamma_{nj} \in \mathbb{R}$ with $\gamma_{nj} \to 1$ as $n \to \infty$ such that the trigonometric polynomials

$$f_n(t) = \sum_1^n \gamma_{nj} a(\rho_j, f)\, e^{i\rho_j t}$$

converge to f uniformly on $\mathbb{R}$, provided $\exp(f) \subset \{\rho_j\}_1^\infty$ (cf. Amerio and Prouse [1] for the proofs).

Recall also that a function $f \in C_u(X)$ is called *asymptotically almost periodic* (a.a.p.) if there is $g \in AP(X)$ such that $|f(t) - g(t)| \to 0$ for $t \to \infty$. Such g is unique since $a(\rho, f) = a(\rho, g)$ for $\rho \in \mathbb{R}$; it is called the a.p.-part of f and will be denoted by f_a. It is not difficult to verify that the space $AAP(X)$ of all a.a.p.-functions is a closed subspace of $C_u(X)$ and decomposes according to

$$AAP(X) = AP(X) \oplus C_0^+(X),$$

where

$$C_0^+(X) = \left\{ f \in C_u(X) : \lim_{t \to \infty} f(t) = 0 \right\}.$$

Thus for any $f \in APP(X)$ we have that $f = f_a + f_0$ with unique $f_a \in AP(X)$ and $f_0 \in C_0^+(X)$. Note that

$$C_l^+(X) = \left\{ f \in C_u(X) : \lim_{t \to \infty} f(t) = f(\infty) \text{ exists} \right\}$$

is a closed subspace of $AAP(X)$.

We are now in position to state our main result on a.p. and a.a.p. solutions.

THEOREM 4. *Let $W = \Lambda(X)$ be admissible for* (1) *and let G denote the corresponding solution operator. Then*

(i) $f \in AP(X)$, $\exp(f) \subset \Lambda$ *imply* $Gf \in AP(X)$, $\exp(Gf) = \exp(f)$ *and*

$$a(\rho, Gf) = H(i\rho)a(\rho, f) \quad \text{for all } \rho \in \mathbb{R}. \tag{13}$$

(ii) $f \in AAP(X)$, $\sigma(f) \subset \Lambda$ *imply* $Gf \in AAP(X)$, $(Gf)_a = Gf_a$, $(Gf)_0 = Gf_0$, *and* (13) holds.

(iii) $f \in C_l^+(X)$, $\sigma(f) \subset \Lambda$, $0 \in \Lambda$ *imply* $Gf \in C_l^+(X)$ *and*

$$(Gf)(\infty) = H(0)f(\infty). \tag{14}$$

It should be mentioned that Theorem 4(i) contains as special cases results on periodic solutions of (1) for periodic f. For instance, if $\Lambda = \{2\pi n/\tau : n \in \mathbb{Z}\}$ and $\Lambda(X)$ is admissible, then each τ-periodic function $f \in C(X)$ satisfies $\exp(f) = \sigma(f) \subset \Lambda$; hence there is precisely one τ-periodic solution u of (1) and, moreover, we have the relation

$$u_n = H(2\pi i n/\tau) f_n, \qquad n \in \mathbb{Z}$$

for the Fourier-coefficients of f and u.

5. Nonresonant equations. The results of §§ 2–4 can be applied to the study of the asymptotic behavior of bounded solutions of (2) and of the resolvent $S(t)$ of (2), provided (1) is *nonresonant*, which means that $W = C_u(X)$ is admissible for (1). Note that in this case Theorems 2 and 3 yield G_0 such that $\tilde{G}_0(\rho) = H(i\rho)$ for all $\rho \in \mathbb{R}$ and the kernel G_0 represents the solution operator by means of (10). More generally, we define the following.

DEFINITION 4. Suppose (1) is nonresonant and let G denote Green's operator for $W = C_u(X)$. An operator-valued function $G_0 : \mathbb{R} \to \mathbb{B}(X)$ is called *solution kernel* for (1) if $G_0(t)$ satisfies

(G1) $G_0(\cdot)x$ is continuous on $\mathbb{R}\setminus\{0\}$ for each $x \in X$;

(G2) $G_0(\cdot)x \in L^1(X) \cap L^\infty(X)$ for each $x \in X$;

(G3) $\tilde{G}_0(\rho) = H(i\rho)$ for each $\rho \in \mathbb{R}$;

(G4) $G_0(0+)x - G_0(0-)x = x$ for each $x \in X$.

Note that solution kernels are unique.

Suppose that (1) is nonresonant and admits the solution kernel $G_0(t)$. Let $S(t)$ be a resolvent for (2) of exponential growth, i.e.,

$$|S(t)| \leqq M e^{\omega t} \quad \text{for all } t \geqq 0 \tag{E}$$

holds with some $M \geqq 1$ and $\omega \in \mathbb{R}$, and suppose that the (complex) *spectrum* of (2)

$$\Sigma_0 = \{\lambda \in \mathbb{C} : \operatorname{Re}\lambda \geqq 0,\ \lambda - A - \widehat{dB}(\lambda) \text{ is not invertible}\}$$

is compact and that

$$|H(\lambda)| \leqq M \quad \text{for } \operatorname{Re}\lambda \geqq 0,\ \operatorname{dist}(\lambda, \Sigma_0) \geqq 1 \tag{15}$$

holds. Since $\hat{S}(\lambda) = H(\lambda)$ the inversion theorem for the Laplace-transform yields for $x \in D(A)$ and some $\omega_1 > \omega$

$$S(t)x = (2\pi i)^{-1} \lim_{N\to\infty} \int_{\omega_1 - Ni}^{\omega_1 + Ni} e^{\lambda t} H(\lambda) x \, d\lambda \quad \text{for } t \neq 0.$$

Thus, shifting the path of integration to the imaginary axis, we obtain

$$S(t)x = (2\pi)^{-1} \lim_{N\to\infty} \int_{-N}^{N} e^{i\rho t} H(i\rho) x \, d\rho + (2\pi i)^{-1} \int_\Gamma e^{\lambda t} H(\lambda) x \, d\lambda$$

where Γ denotes some Jordan curve contained in $\{\operatorname{Re}\lambda > 0\}$ and surrounding Σ_0. The inversion theorem for the Fourier-transform therefore yields

$$S(t)x = G_0(t)x + S_0(t)x \tag{16}$$

with

$$S_0(t) = (2\pi i)^{-1} \int_\Gamma e^{\lambda t} H(\lambda)\, d\lambda, \qquad t \in \mathbb{R}. \tag{17}$$

Note that $S_0 \in C^\infty(\mathbb{R}, \mathbb{B}(X, Y))$ and

$$S_0'(t) = AS_0(t) + \int_0^\infty dB(\tau) S_0(t-\tau) \quad \text{for } t \in \mathbb{R}.$$

In the case where $\Sigma_0 = \{\lambda_1, \cdots, \lambda_n\}$ is finite and each λ_j is a pole of finite multiplicity m_j, $S_0(t)$ can be evaluated by means of the calculus of residues to the result

$$S_0(t) = \sum_{j=1}^{n} e^{\lambda_j t} \sum_{m=1}^{m_j} H_{j,m} t^{m-1}/(m-1)!, \qquad t \in \mathbb{R} \tag{18}$$

for some $H_{j,m} \in \mathbb{B}(X, Y)$, i.e., $S_0(t)$ becomes a generalized exponential polynomial.

Property (G3) and the variation of parameters formula (3) now yield the following result.

THEOREM 5. *Suppose* (1) *is nonresonant and admits a solution kernel* $G_0(t)$. *Let* $S(t)$ *be a resolvent for* (2) *with* (E) *and suppose* Σ_0 *is compact and that* (15) *holds. Then*

$$S(t) = G_0(t) + S_0(t) \quad \textit{for all } t \in \mathbb{R}\setminus\{0\} \tag{16}$$

where $S_0 \in C^\infty(\mathbb{R}, \mathbb{B}(X, Y))$ *is given by* (17). *Any solution* $u \in C_u(\mathbb{R}_+, X)$ *of* (2) *with* $g = f$ *satisfies*

$$u(t) - (Gf)(t) \to 0 \quad \textit{as } t \to \infty;$$

in particular

(i) u *is a.a.p. if* $f \in AAP(X)$, *and* $a(\rho, u) = H(i\rho)a(\rho, f)$ *for all* $\rho \in \mathbb{R}$,

(ii) $u \in C_l^+(\mathbb{R}_+, X)$ *if* $f \in C_l^+(X)$, *and* $\lim_{t\to\infty} u(t) = H(0)f(\infty)$.

Combining Theorem 1 and Theorem 4 we obtain the following generalization of results of Friedman and Shinbrot [5] and Miller and Wheeler [12] concerning integrable resolvents of (2).

COROLLARY 1. *Suppose* $\Sigma_0 = \varnothing$. *Then*

(i) *if the assumptions of Theorem* 2 *are satisfied, the resolvent* $S(t)$ *of* (2) *exists and belongs to* $L^1(\mathbb{R}_+, \mathbb{B}(X)) \cap C_0(\mathbb{R}_+, \mathbb{B}(X))$;

(ii) *if the assumptions of Theorem* 3 *are satisfied, the resolvent* $S(t)$ *of* (2) *exists and* $S(t)x$ *belongs to* $L^1(\mathbb{R}_+, X) \cap C_0(\mathbb{R}_+, X)$ *for each* $x \in X$.

In either case each solution u *of* (2) *with* $g = f \in C_u(X)$ *asymptotically behaves like the solution* Gf *of* (1), *i.e.,* $u(t) - Gf(t) \to 0$ *as* $t \to \infty$.

6. Applications. (a) Consider the heat equation in materials with memory

$$\begin{aligned} &u_t(t, x) = \Delta u(t, x) + \int_0^\infty b(\tau)\Delta u(t-\tau)\, d\tau + f(t, x), \qquad t \in \mathbb{R}, \quad x \in \Omega, \\ &u(t, x) = 0, \qquad t \in \mathbb{R}, \quad x \in \partial\Omega, \end{aligned} \tag{19}$$

where $\Omega \in \mathbb{R}^N$ denotes a bounded domain with smooth boundary $\partial\Omega$ and $b \in L^1(\mathbb{R}_+)$ (cf., e.g., Coleman and Gurtin [2]).

We let $X = L^2(\Omega)$, $A = \Delta$ with $D(A) = W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)$. A is selfadjoint and negative definite hence analytic, and $\sigma(A) = \{\mu_j\}_1^\infty \subset (-\infty, 0)$. Formula (19) then becomes

$$u' = Au + b * Au + f \tag{20}$$

and clearly the assumptions of Theorem 2 are satisfied. To determine the spectra Σ_0 and Λ_0 of (20) note that $\lambda - A - \widehat{dB}(\lambda) = \lambda - (1 + \hat{b}(\lambda))A$ with domain $D(A)$ is invertible iff $1 + \hat{b}(\lambda) \neq 0$ and $\lambda(1 + \hat{b}(\lambda))^{-1} \notin \sigma(A)$, i.e., $\lambda \neq \mu_j(1 + \hat{b}(\lambda))$ for all j. With

$$\begin{aligned} &\chi_j(\lambda) = 1 + \hat{b}(\lambda) - \lambda/\mu_j, \qquad j \in \mathbb{N}, \\ &\chi_\infty(\lambda) = 1 + \hat{b}(\lambda) \end{aligned} \tag{21}$$

we therefore have

$$\Sigma_0 = \{\lambda \in \mathbb{C}: \operatorname{Re} \lambda \geqq 0, \chi_j(\lambda) = 0 \text{ for some } j \in \mathbb{N} \cup \{\infty\}\}$$

and

$$\Lambda_0 = \{\rho \in \mathbb{R}: \chi_j(i\rho) = 0 \text{ for some } j \in \mathbb{N} \cup \{\infty\}\};$$

Λ_0 and Σ_0 are compact, $\Sigma_0 \cap \{\operatorname{Re} \lambda > 0\}$ is at most countable, Λ_0 has Lebesgue-measure zero, and $\lambda_0 \in \Sigma_0$ with $\operatorname{Re} \lambda_0 > 0$ is a cluster point of Σ_0 iff $\chi_\infty(\lambda_0) = 0$. Theorem 2 applies, hence $\Lambda(X)$ is admissible for (20) iff $\Lambda \cap \Lambda_0 = \varnothing$, in particular, (20) is nonresonant iff $\chi_j(i\rho) \neq 0$ for all $\rho \in \mathbb{R}$, $j \in \mathbb{N} \cup \{\infty\}$. In this case there is a solution kernel $G_0 \in L^1(\mathbb{B}(X))$ for (20) and the resolvent $S(t)$ admits the decomposition $S(t) = G_0(t) + S_0(t)$ by Theorem 5. If we also have $\chi_\infty(\lambda) \neq 0$ for $\operatorname{Re} \lambda \geqq 0$, then Σ_0 is finite and

$$S_0(t) = \sum_{j=1}^{n} e^{\lambda_j t} \sum_{m=1}^{m_j} H_{j,m} t^{m-1}/(m-1)!$$

where m_j denotes the multiplicity of $\lambda_j \in \Sigma_0$. Finally, by Theorem 5 we have $S \in L^1(\mathbb{B}(X))$ iff $\chi_j(\lambda) \neq 0$ for all $\operatorname{Re} \lambda \geqq 0$ and $j \in \mathbb{N} \cup \{\infty\}$. This generalizes the results of Friedman and Shinbrot [5] and Miller and Wheeler [12].

(b) The following system of equations arise as linearizations of nonlinear equations at equilibrium points in population dynamics (cf. Cushing [3]):

$$\begin{aligned} &u_{jt}(t, x) = d_j \Delta u_j(t, x) + \sum_{k=1}^{n} \int_0^\infty db_{jk}(\tau) u_k(t - \tau, x), \qquad t \in \mathbb{R}, \quad x \in \Omega, \\ &d_j \frac{\partial u_j}{\partial \nu}(t, x) = 0, \quad x \in \partial\Omega, \quad j = 1, \cdots, n. \end{aligned} \tag{22}$$

Here $\Omega \in \mathbb{R}^N$ denotes again a bounded domain with smooth boundary and $\nu(x)$ the outer normal at $x \in \partial\Omega$. The solvability behavior of (22) is important for stability and bifurcation considerations for the nonlinear system.

Let $X = [L^2(\Omega)]^n$, $A = D \cdot \Delta = (\operatorname{diag} d_j)\Delta$, $d_j \geqq 0$ for all j, with domain $D(A) = \{u \in X: d_j u_j \in W^{2,2}(\Omega), d_j \partial u_j/\partial \nu = 0 \text{ on } \partial\Omega\}$ and let $B(t)$ be defined according to

$$(B(t)u)_j(x) = \sum_{k=1}^{n} b_{jk}(t) u_k(x), \qquad u \in X,$$

where $b_{jk} \in BV(\mathbb{R}_+)$. Then A is analytic and Theorem 2 is applicable. It therefore remains to compute Σ_0 and Λ_0. For this purpose let $\mu_0 = 0 > \mu_1 \geqq \mu_2 \geqq \cdots$ denote the eigenvalues of the Laplacian with Neumann boundary condition and e_j the corresponding eigenvectors. A simple calculation then yields

$$\Sigma_0 = \{\lambda \in \mathbb{C}: \operatorname{Re} \lambda \geqq 0, \chi_m(\lambda) = 0 \text{ for some } m \in \mathbb{N}_0\}$$

and

$$\Lambda_0 = \{\rho \in \mathbb{R}: \chi_m(i\rho) = 0 \text{ for some } m \in \mathbb{N}_0\}$$

where

$$\chi_m(\lambda) = \det(\lambda - \mu_m D - \widehat{dB}(\lambda)), \qquad m \in \mathbb{N}_0.$$

Note that $\chi_m(\lambda) \neq 0$ for $m \geqq m_0$, $\operatorname{Re}\lambda \geqq 0$; hence $\Sigma_0 \setminus \Lambda_0$ is countable and discrete. Λ_0 has Lebesgue-measure zero.

(c) The following problem arises in the theory of viscoelasticity (cf., e.g., Dafermos [4] and the references given there):

$$\begin{aligned} &v_{tt}(t,x) + \gamma v_t(t,x) = \Delta v(t,x) - (b * \Delta v)(t,x) + g(t,x), \qquad t \in \mathbb{R}, \quad x \in \Omega, \\ &\frac{\partial v}{\partial \nu}(t,x) + c(x)v(t,x) = 0, \qquad x \in \partial\Omega, \quad t \in \mathbb{R}, \end{aligned} \tag{23}$$

where Ω and ν are as before, $c \in C(\partial\Omega, \mathbb{R}_+)$, $c \not\equiv 0$ and $b \in W^{1,1}(\mathbb{R}_+, \mathbb{R})$ such that $\int_0^\infty t|b(t)|\,dt + \int_0^\infty t|b'(t)|\,dt < \infty$.

Let $H_0 = L^2(\Omega)$, $H_1 = W^{1,2}(\Omega)$, $H_2 = W^{2,2}(\Omega)$ and $\mathcal{A} = \Delta$ with domain $D(\mathcal{A}) = \{v \in H_2: \partial v/\partial\nu + cv = 0 \text{ on } \partial\Omega\}$; $\mathcal{A}$ is selfadjoint and negative definite in H_0 and $\sigma(A) = \{\mu_j\}_1^\infty \subset (-\infty, 0)$. Formula (23) then becomes the abstract second order equation

$$v'' + \gamma v' = \mathcal{A}v - b * \mathcal{A}v + g, \tag{24}$$

which can be reduced to (1) as usual. Let $X = H_1 \times H_0$, $u = (v, v')$, $f = (0, g)$ and

$$A = \begin{pmatrix} 0 & I \\ \mathcal{A} & -\gamma I \end{pmatrix}, \qquad B'(t) = \begin{pmatrix} 0 & 0 \\ -b(t)\mathcal{A} & 0 \end{pmatrix}$$

where $D(A) = D(\mathcal{A}) \times H_1$. It is well known that A generates a C_0-group in X and $B(t)$ meets the assumptions of Theorem 3. To compute the spectra Σ_0 and Λ_0 we note that

$$\lambda - A - \widehat{B'}(\lambda) = \begin{pmatrix} \lambda & -I \\ (\hat{b}(\lambda) - 1)\mathcal{A} & (\lambda + \gamma)I \end{pmatrix}$$

is invertible in X iff $(\lambda + \gamma)\lambda - (1 - \hat{b}(\lambda))\mathcal{A}$ is invertible in H_0; hence we obtain

$$\Sigma_0 = \{\lambda \in \mathbb{C}: \operatorname{Re}\lambda \geqq 0 \text{ and } \chi_m(\lambda) = 0 \text{ for some } m \in \mathbb{N} \cup \{\infty\}\}$$

and

$$\begin{aligned} &\Lambda_0 = \{\rho \in \mathbb{R}: \chi_m(i\rho) = 0 \text{ for some } m \in \mathbb{N} \cup \{\infty\}\}, \\ &\chi_m(\lambda) = 1 - \hat{b}(\lambda) - \lambda(\lambda + \gamma)/\mu_m, \qquad m \in \mathbb{N} \end{aligned}$$

where

$$\chi_\infty(\lambda) = 1 - \hat{b}(\lambda).$$

The representation

$$H(\lambda) = F(\lambda)\begin{pmatrix} (\lambda + \gamma)I & I \\ (1 - \hat{b}(\lambda))\mathcal{A} & \lambda I \end{pmatrix}, \qquad F(\lambda) = (\lambda(\lambda + \gamma) - (1 - \hat{b}(\lambda))\mathcal{A})^{-1}$$

yields the estimates

$$|H(i\rho)|_X \leqq M|\rho|\,|F(i\rho)|_{H_0}, \qquad |AH(i\rho)|_X \leqq M|\rho|^2|F(i\rho)|_{H_0}$$

for $|\rho| \to \infty$, $\rho \in \Lambda_0$, since it is well known that $H_1 = D((-\mathscr{A})^{1/2})$ holds (cf. Tanabe [15, Thm. 2.23]). On the other hand, we have

$$|F(i\rho)|_{H_0} \leqq [\inf_m |-\rho^2 + i\gamma\rho - (1-\hat{b}(i\rho))\mu_m|]^{-1}$$

$$\leqq |\rho|^{-1} \cdot |1-\hat{b}(i\rho)| / |-\rho \operatorname{Im} \hat{b}(i\rho) + \gamma \cdot (1 - \operatorname{Re} \hat{b}(i\rho))|.$$

The relation

$$-\rho \operatorname{Im} \hat{b}(i\rho) = \rho \int_0^\infty b(t) \sin \rho t \, dt = b(0) + \int_0^\infty b'(t) \cos \rho t \, dt = b(0) + \operatorname{Re} \widehat{b'}(i\rho)$$

yields

$$-\rho \operatorname{Im} \hat{b}(i\rho) + \gamma(1 - \operatorname{Re} \hat{b}(i\rho)) \to b(0) + \gamma \quad \text{as } |\rho| \to \infty;$$

hence we obtain

$$|F(i\rho)|_{H_0} \leqq C|\rho|^{-1} \cdot |b(0) + \gamma|^{-1} \quad \text{for } |\rho| \geqq \rho_0.$$

Therefore the estimates required in Theorem 3 are satisfied if $b(0) + \gamma \neq 0$, in particular Λ_0 is compact, and Theorem 3 shows that $\Lambda(X)$ is admissible iff $\Lambda \cap \Lambda_0 = \varnothing$, and (24) is nonresonant iff $\Lambda_0 = \varnothing$. It is easy to see that the latter is the case if, e.g., $\gamma \geqq 0$, $b(t) \geqq 0$ nonincreasing and $\int_0^\infty b(t)\, dt \neq 1$ holds. A calculation similar to the one above shows also that Σ_0 is compact and (13) holds, in the case $\gamma + b(0) > 0$, hence Theorem 5 yields the decomposition

$$S(t) = G_0(t) + S_0(t)$$

for the resolvent $S(t)$, where $S_0 \in C_b^\infty(\mathbb{R}, \mathbb{B}(X, Y)) \cap L^1(\mathbb{R}_-, \mathbb{B}(X, Y))$ and $G_0(\cdot)x$ is in $L^1(X)$ for each $x \in X$. Finally, for $\gamma \geqq 0$, $b(t) > 0$ nonincreasing, $b(0) > 0$ and $\int_0^\infty b(t)\, dt < 1$ we even obtain $\Sigma_0 = \varnothing$, hence $S(\cdot)x \in L^1(\mathbb{R}_+, X)$ for all $x \in X$; this result has been proved by Dafermos [4], assuming additionally that b is convex.

7. Proofs.

(a) *Proof of Lemma* 1.

Since $\tilde{\varphi}$ has compact support and $\Lambda_0 \cap \operatorname{supp} \tilde{\varphi} = \varnothing$, $H(i\rho)$ is bounded in $\mathbb{B}(X, Y)$ on $\operatorname{supp} \tilde{\varphi}$, and therefore G_φ defined by

$$G_\varphi(t) = (2\pi)^{-1} \int_{-\infty}^{\infty} H(i\rho)\tilde{\varphi}(\rho)\, e^{i\rho t}\, d\rho \tag{25}$$

belongs to $C_b^\infty(\mathbb{B}(X, Y))$.

(i) To obtain $G_\varphi \in L^1(\mathbb{B}(X, Y))$ suppose first that $\int_0^\infty (1+t^2)\, db(t) < \infty$ where $b(t) = \operatorname{Var} B|_0^t$ denotes the variation of B in $\mathbb{B}(Y, X)$. Then $\widehat{dB}(i\rho)$ is twice continuously differentiable, and so is $\Phi(\rho) = H(i\rho)\tilde{\varphi}(\rho)$ and $\Phi''(\rho)$ is bounded on $\mathbb{R}$. Partial integration in (25) now yields

$$G_\varphi(t) = -(2\pi)^{-1} t^{-2} \int_{-\infty}^{\infty} \Phi''(\rho)\, e^{i\rho t}\, ds \quad \text{for } t \neq 0$$

and so we obtain an estimate $|G(t)|_{X,Y} \leqq C(1+t^2)^{-1}$ for $t \in \mathbb{R}$, i.e., $G_\varphi \in L^1(\mathbb{B}(X, Y))$ holds.

(ii) For the general case we let $B_m(t) = B(t)$ for $t \leqq m$ and $B_m(t) = B(m)$ for $t \geqq m$; then $\widehat{dB}_m(i\rho)$ has moments of any order and $\widehat{dB}_m(\lambda) \to \widehat{dB}(\lambda)$ for $m \to \infty$ in $\mathbb{B}(Y, X))$ uniformly on $\operatorname{Re} \lambda \geqq 0$. Therefore the relation

$$i\rho - A - \widehat{dB}_m = (i\rho - A - \widehat{dB})[I - H(i\rho)(\widehat{dB}_m - \widehat{dB})] \tag{26}$$

shows that $H_m(i\rho) = (i\rho - A - \widehat{dB}_m)^{-1}$ exists on a fixed neighborhood N of supp $\tilde{\varphi}$, in the case where m is chosen large enough; note that $H(i\rho)$ is bounded in $\mathbb{B}(X, Y)$ on N. Equation (26) yields

$$H(i\rho) = H_m(i\rho) + H(i\rho)[\widehat{dB}(i\rho) - \widehat{dB}_m(i\rho)]H_m(i\rho);$$

hence

$$H(i\rho)\tilde{\varphi}(\rho) = H_m(i\rho)\tilde{\varphi}(\rho) + H(i\rho)\tilde{\varphi}(\rho)[\widehat{dB}(i\rho) - \widehat{dB}_m(i\rho)]H_m(i\rho)\tilde{\varphi}_1(\rho), \tag{27}$$

where $\tilde{\varphi}_1 \in C_0^\infty$ is such that $\tilde{\varphi}_1(\rho) \equiv 1$ on supp $\tilde{\varphi}$, $0 \leqq \tilde{\varphi} \leqq 1$, and supp $\tilde{\varphi}_1 \subset N$. By step (i) of this proof, there are G_m, $K_m \in C_b^\infty(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ such that $\tilde{G}_m(\rho) = H_m(i\rho)\tilde{\varphi}(\rho)$ and $\tilde{K}_m(\rho) = H_m(i\rho)\tilde{\varphi}_1(\rho)$ (choose N such that $N \cap \Lambda_0 = \varnothing$); let $K = (\widehat{dB} - \widehat{dB}_m) * K_m$ and observe that $K \in L^1(\mathbb{B}(X))$ holds. Now we have

$$\begin{aligned} I - \tilde{K} &= I - (\widehat{dB} - \widehat{dB}_m)\tilde{K}_m = I - (\widehat{dB} - \widehat{dB}_m)H_m\tilde{\varphi}_1 \\ &= (i\rho - A - \widehat{dB}\tilde{\varphi}_1 - \widehat{dB}_m(1 - \tilde{\varphi}_1))H_m \\ &= [I - (1 - \tilde{\varphi}_1)(\widehat{dB}_m - \widehat{dB})H](i\rho - A - \widehat{dB})H_m, \end{aligned}$$

i.e., $I - K(\rho)$ is invertible on $\mathbb{R}$, and Lemma 2 yields a solution $L \in L^1(\mathbb{B}(X))$ of (11). Therefore the equation $G_\varphi = G_m + G_\varphi * K$ whose Fourier-transform is (27) has the solution $G_\varphi = G_m + G_m * L$ which belongs to $L^1(\mathbb{B}(X, Y))$. Uniqueness of the Fourier-transform now gives the assertion of Lemma 1. □

(b) *Proof of Proposition* 1.

(i) Since for any $y \in X$, $\rho \in \mathbb{R}$, the function $f(t) = y\, e^{i\rho t}$ belongs to $C_u(X)$ and satisfies $\sigma(f) = \{\rho\}$, we have $f \in W_0 = \Lambda(X) \cap C_u^1(X)$ in the case where $\rho \in \Lambda$. Now, if $\Lambda(X)$ is admissible for (1), there is a solution $u = Gf$ of (1) and we obtain

$$u' = (Gf)' = Gf' = i\rho Gf = i\rho u;$$

hence $u(t) = x\, e^{i\rho t}$ for some $x \in D(A)$, and (1) implies $(i\rho - A - \widehat{dB}(i\rho))x = y$, i.e., $i\rho - A - \widehat{dB}(i\rho)$ is surjective. On the other hand, we have for some $M > 0$

$$|x| = |u|_0 = |Gf|_0 \leqq M|f|_0 = M|(i\rho - A - \widehat{dB}(i\rho))x|;$$

i.e., $i\rho - A - \widehat{dB}(i\rho)$ is bounded from below uniformly on Λ, and this yields $|H(i\rho)| \leqq M$ in $\mathbb{B}(X)$ on Λ, in particular $\Lambda \cap \Lambda_0 = \varnothing$. Since G is also bounded as a map from $\Lambda(X) \cap C_u^1(X)$ to $C_u(Y)$ by the closed graph theorem, we finally obtain

$$|AH(i\rho)y| = |Ax| \leqq |Gf|_{Y,0} \leqq M_1(|f'|_0 + |f|_0) \leqq M_1(1 + |\rho|)|y|;$$

i.e., the second estimate of Proposition 1(i) holds.

(ii) For the proof of (ii) and (iii) of Proposition 1 we shall need some properties of the spectrum of bounded functions which are collected in the following.

PROPOSITION 2. *Let $f, g \in C_b(X)$ and $K \in BV(\mathbb{B}(X, Z))$. Then*

(i) *$\sigma(f)$ is closed and $\neq \varnothing$ in case $f \neq 0$;*

(ii) *$\sigma(f + g) \subset \sigma(f) \cup \sigma(g)$;*

(iii) *$\sigma(f') \subset \sigma(f)$ for each $f \in C_b^1(X)$;*

(iv) *$\sigma(dK * f) \subset \sigma(f)$;*

(v) *$\Lambda \subset \mathbb{R}$ closed, $(f_n) \subset C_b(X)$ uniformly bounded, $\sigma(f_n) \subset \Lambda$ for each n, and $f_n \to f$ uniformly on compact sets imply $\sigma(f) \subset \Lambda$; i.e., $\sigma(\cdot)$ is lower semicontinuous;*

(vi) *$\sigma(f) = \operatorname{supp} \widehat{D_f}$, where $\widehat{D_f}$ denotes the Fourier-transform of the distribution D_f induced by f;*

(vii) *$\sigma_X(f) = \sigma_Y(f)$ for each $f \in C_b(Y)$, where $Y \to X$;*

(viii) *$\varphi * f = f$ for each $\varphi \in L^1$ such that $\tilde{\varphi} \equiv 1$ on $\sigma(f)$.*

Proofs for these properties can be found in Katznelson [11] in the case of scalar functions or Prüss [13] for the general case.

Now, let $\tilde{\varphi} \in C_0^\infty$ be such that $\operatorname{supp} \tilde{\varphi} \cap \sigma(f) = \emptyset$. Since $\sigma(\varphi * f) \subset \sigma(\varphi) \cap \sigma(f) = \operatorname{supp} \tilde{\varphi} \cap \sigma(f) = \emptyset$ we have $\varphi * f = 0$ and therefore

$$\langle \widetilde{D_{Gf}}, \tilde{\varphi} \rangle = \langle Gf, \tilde{\tilde{\varphi}} \rangle = (\varphi * Gf)(0) = G(\varphi * f)(0) = 0.$$

Hence we obtain $\sigma(Gf) = \operatorname{supp} \widetilde{D_{Gf}} \subset \sigma(f)$ by definition of the support of a distribution.

Conversely, suppose $f \in \Lambda(X) \cap C_u^1(X)$. Then by (1) we get

$$\sigma(f) = \sigma(u' - Au - dB * u) \subset \sigma_Y(u) = \sigma(u)$$

where $u = Gf$. For the general case let φ_n denote a Dirac-sequence, i.e., $\varphi_n \in C_0^\infty$, $\operatorname{supp} \varphi_n \subset (-n^{-1}, n^{-1})$, $\varphi_n \geqq 0$ and $\int_{-\infty}^{\infty} \varphi_n(t)\, dt = 1$ for each $n \in \mathbb{N}$. Then $f_n = \varphi_n * f$ is uniformly bounded, converges to f as $n \to \infty$ in $C_u(X)$ and $f_n \in \Lambda(X) \cap C_u^1(X)$ for each n. We therefore have

$$\sigma(f_n) \subset \sigma(Gf_n) = \sigma(G(\varphi_n * f)) = \sigma(\varphi_n * Gf) \subset \sigma(Gf);$$

hence $\sigma(f) \subset \sigma(Gf)$ by (v) and (i) of Proposition 2.

(iii) Suppose $u \in C_u(Y) \cap C_u^1(X)$ is a solution of equation (1) with $f = 0$ on $\mathbb{R}$, and let $\tilde{\varphi} \in C_0^\infty$ be such that $\Lambda_0 \cap \operatorname{supp} \tilde{\varphi} = \emptyset$. By Lemma 1, there is a solution $G_\varphi \in C_u^\infty(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ of (10). This implies

$$G_\varphi * Au + G_\varphi * dB * u = G_\varphi * u' = G'_\varphi * u = G_\varphi * Au + G_\varphi * dB * u + \varphi * u,$$

i.e., $\varphi * u \equiv 0$; hence

$$\langle \widetilde{D_u}, \tilde{\varphi} \rangle = \langle u, \tilde{\tilde{\varphi}} \rangle = (\varphi * u)(0) = 0,$$

which means $\sigma(u) = \operatorname{supp} \widetilde{D_u} \subset \Lambda_0$. □

(c) *Proof of Theorem* 1.

(i) Suppose $u_1, u_2 \in \Lambda(X) \cap C_u^1(X)$ are two solutions of (1) for the same $f \in \Lambda(X)$. Then $u = u_1 - u_2 \in \Lambda(X) \cap C_u^1(X)$ is a solution of the homogeneous equation (1). Hence by Proposition 1(iii) we have $\sigma(u) \subset \Lambda_0 \cap \Lambda = \emptyset$, i.e., $u = 0$ by Proposition 2(i). This shows that solutions are unique.

(ii) To obtain the existence of a solution for f from a dense subset of $\Lambda(X)$, let $\tilde{\varphi} \in C_0^\infty$ be such that $0 \leqq \tilde{\varphi} \leqq 1$, $\tilde{\varphi}(\rho) = 1$ for $|\rho| \leqq 1$, $\tilde{\varphi}(\rho) = 0$ for $|\rho| \geqq 2$ and consider $\varphi_n(t) = n\varphi(nt)$. Then $\{\varphi_n\}$ is an approximation of the identity, i.e., $\varphi_n * f \to f$ as $n \to \infty$ in $C_u(X)$. Hence it is sufficient to solve (1) for $\varphi_n * f$, $f \in \Lambda(X)$, $n \in \mathbb{N}$. But since $\sigma(\varphi_n * f) \subset \sigma(f) \cap \sigma(\varphi_n) \subset \Lambda \cap \{\rho \in \mathbb{R}: |\rho| \leqq 2n\}$ holds we only need to solve (1) for $f \in \Lambda(X)$ with $\sigma(f)$ compact; i.e., we may assume that $\Lambda \in \mathbb{R}$ is compact.

(iii) Suppose $\Lambda \subset \mathbb{R}$ is compact. Then we choose a cutoff function $\tilde{\varphi}_0 \in C_0^\infty$ such that $\tilde{\varphi}_0 \equiv 1$ for $\operatorname{dist}(\rho, \Lambda) \leqq \varepsilon$, $\tilde{\varphi}_0 \equiv 0$ for $\operatorname{dist}(\rho, \Lambda) \geqq 2\varepsilon$, $0 \leqq \tilde{\varphi}_0 \leqq 1$, where $\operatorname{dist}(\Lambda, \Lambda_0) \geqq 3\varepsilon$, and let $G_\Lambda \in C_b^\infty(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ denote the solution of (10) from Lemma 1. Then $u = G_\Lambda * f$ is a solution of (1) with inhomogeneity $\varphi_0 * f$ for each $f \in C_u(X)$ and we have $|u|_0 \leqq |G_\Lambda|_{L^1} |f|_0$. But for $f \in \Lambda(X)$ we obtain $\varphi_0 * f = f$ by Proposition 2(viii) and so u is already a solution of (1). This shows that $\Lambda(X)$ is admissible and G is represented by (12). □

(d) *Proof of Theorem* 4.

(i) Suppose $\Lambda(X)$ is admissible. Then if $\rho \in \Lambda$, as in (b) we obtain $G(a\, e^{i\rho t}) = H(i\rho) a\, e^{i\rho t}$ for each $a \in X$. Hence if $f(t) = \sum_{n=1}^N a_n\, e^{i\rho_n t}$ is a trigonometric polynomial with $\exp(f) = \{\rho_n\}_1^N \subset \Lambda$ we obtain $(Gf)(t) = \sum_{n=1}^N H(i\rho_n)\, a_n\, e^{i\rho_n t}$, i.e., (i) of Theorem 4 holds for trigonometric polynomials. The general case then follows by Bochner's approximation theorem mentioned above since G is bounded and the Bohr-transform is continuous.

(ii) If $f=f_a+f_0\in AAP(X)\cap\Lambda(X)$, then $T_nf\to f_a$ as $n\to\infty$ uniformly on compact subsets; hence by Proposition 2(v) and from $\sigma(T_nf)=\sigma(f)\subset\Lambda$ we obtain $\sigma(f_a)\subset\Lambda$ and so $\sigma(f_0)=\sigma(f-f_a)\subset\sigma(f)\cup\sigma(f_a)\subset\Lambda$, also. Since $\Lambda(X)$ is admissible we therefore obtain $Gf=Gf_a+Gf_0$ and $Gf_a\in AP(X)$ by step (i) of this proof; hence it remains to show that $Gf_0\in C_0^+(X)$ holds. As in (c), step (ii) it suffices to consider $f_0\in C_0^+(X)\cap\Lambda(X)$ with $\sigma(f_0)$ compact, i.e., w.l.o.g. Λ compact. But in this case Lemma 1 yields a kernel $G_\Lambda\in C_b^\infty(\mathbb{B}(X,Y))\cap L^1(\mathbb{B}(X,Y))$ such that $Gf=G_\Lambda*f$ for all $f\in\Lambda(X)$, and since $G_\Lambda\in L^1(\mathbb{B}(X,Y))$ we obtain $Gf\in C_0^+(X)$ for each $f\in C_0^+(X)\cap\Lambda(X)$.

(iii) This is a consequence of (ii) with $f_a(t)\equiv f(\infty)$. □

(e) *Proof of Theorem* 2.

(i) Since A is analytic, $(i\rho-A)^{-1}$ exists say for $|\rho|\geqq N_0$ and satisfies $|(i\rho-A)^{-1}|\leqq M/|\rho|$ for $|\rho|\geqq N_0$. The relation

$$i\rho-A-\widehat{dB}(i\rho)=(I-\widehat{dB}(i\rho)(i\rho-A)^{-1})(i\rho-A),\qquad |\rho|\geqq N_0,$$

then shows that $\Lambda_0\subset(-N,N)$ holds for some $N>0$. In fact, we have $B_1'\in L^1(\mathbb{B}(Y,X))$; hence $B_1'(i\rho)\to 0$ as $|\rho|\to\infty$, by the Riemann-Lebesgue Lemma, and so

$$|\widehat{B_1'}(i\rho)(i\rho-A)^{-1}|_{X,X}\leqq|\widehat{B_1'}(i\rho)|_{Y,X}\cdot|(i\rho-A)^{-1}|_{X,Y}\to 0\quad\text{as } |\rho|\to\infty.$$

Similarly,

$$|\widehat{dB_2}(i\rho)(i\rho-A)^{-1}|_{X,X}\leqq|\widehat{dB_2}(i\rho)|_{Y^\alpha,X}\cdot|(i\rho-A)^{-1}|_{X,Y^\alpha}\leqq C|\rho|^{\alpha-1}\quad\text{as } |\rho|\to\infty,$$

and finally

$$|\widehat{dB_3}(i\rho)(i\rho-A)^{-1}|_{X,X}\leqq|\widehat{dB_3}(i\rho)|_{Y,X}\cdot|(i\rho-A)^{-1}|_{X,Y}\leqq 1/4\quad\text{for } |\rho|\geqq N$$

since B_3 has sufficiently small variation by assumption. Thus w.l.o.g. we let $\Lambda\supset\{\rho\in\mathbb{R}\colon|\rho|\geqq N\}$.

Let $\tilde\varphi_0\in C^\infty$ be a cutoff function for Λ, i.e., $\tilde\varphi_0(\rho)=1$ for dist $(\rho,\Lambda)\leqq\varepsilon$, $\tilde\varphi_0(\rho)=0$ for dist $(\rho,\Lambda)\geqq 2\varepsilon$, $0\leqq\tilde\varphi_0\leqq 1$, where $3\varepsilon\leqq$ dist (Λ,Λ_0). Let $\tilde\varphi_1\in C^\infty$ denote another cutoff function such that $\tilde\varphi_1(\rho)=1$ for $|\rho|\geqq N+1$ and $\tilde\varphi_1(\rho)=0$ for $|\rho|\leqq N$, $0\leqq\tilde\varphi_1\leqq 1$. Then

$$G_1(t)=(2\pi)^{-1}\int_{-\infty}^{\infty}(1-\tilde\varphi_1(\rho))\tilde\varphi_0(\rho)H(i\rho)\,e^{i\rho t}\,d\rho$$

belongs to $C_b^\infty(\mathbb{B}(X,Y))\cap L^1(\mathbb{B}(X,Y))$ by Lemma 1 and so it remains to study

$$G_2(t)=(2\pi)^{-1}\int_{-\infty}^{\infty}\tilde\varphi_1(\rho)H(i\rho)\,e^{i\rho t}\,d\rho;$$

note that $\tilde\varphi_0(\rho)=1$ on supp $\tilde\varphi_1$.

(ii) Let $k_N(t)=2(\pi Nt^2)^{-1}\sin^2(Nt/2)$ denote Fejer's kernel, and let $B_N(t)=B_1(t)-k_N*B_1(t)+B_2(t)+B_3(t)$; note that $\widetilde{(k_N*B_1')}(\rho)=\widehat{k_N}(\rho)\cdot\widehat{B_1'}(i\rho)=0$ for $|\rho|\geqq N$, since $\widehat{k_N}(\rho)=(1-|\rho|/N)_+$ holds, i.e., $\widehat{dB_N}(\rho)=\widehat{dB}(i\rho)$ for $|\rho|\geqq N$. Next we choose $\omega>\omega_0$ sufficiently large and the equivalent norm $\|x\|_Y=|(\omega-A)x|$ on Y. Then the variation of B_N over $\mathbb{R}$ can be made arbitrarily small, since

$$B_1'-k_N*B_1'\to 0\quad\text{in } L^1(\mathbb{B}(Y,X))\quad\text{for } N\to\infty$$

and

$$\operatorname{Var}_{Y,X}B_2|_0^\infty=\operatorname{Var}_{Y^\alpha,X}B_2|_0^\infty\cdot|J|_{Y,Y^\alpha}\leqq\operatorname{Var}_{Y^\alpha,X}B_2|_0^\infty\cdot C\omega^{\alpha-1}\to 0$$

for $\omega\to\infty$, where $J\colon Y\to Y^\alpha$ denotes the natural embedding, and the variation of B_3 is arbitrarily small by assumption. The proof of Theorem 1 in [14] now shows that there is $G_{\omega,N}\in L^1(\mathbb{B}(X))\cap L^\infty(\mathbb{B}(X))$, strongly continuous for $t\neq 0$, satisfying the jump

relation $G_{\omega,N}(0+)-G_{\omega,N}(0-)=I$, such that $G_{\omega,N}(\rho)=(i\rho+\omega-A-\widehat{dB_N}(\rho))^{-1}$ holds. Note that Theorem 1 of [14] does not apply directly since the support of B_N is generally not contained in $\mathbb{R}_+$, however, the method of proof still works for this more general case. Thus the solution $H_{\omega,N}(\rho)=(i\rho+\omega-A-\widehat{dB_N}(\rho))^{-1}$ of the equation

$$H_{\omega,N}(\rho)=(i\rho+\omega-A)^{-1}+H_{\omega,N}(\rho)\widehat{dB_N}(\rho)(i\rho+\omega-A)^{-1} \tag{28}$$

is the Fourier-transform of $G_{\omega,N}\in L^1(\mathbb{B}(X))\cap L^\infty(\mathbb{B}(X))$.

(iii) Since $1-\tilde{\varphi}_1=\tilde{\Psi}_1$ is the Fourier-transform of some function $\Psi_1\in L^1$ we now also have that

$$\tilde{\varphi}_1 H_\omega=\tilde{\varphi}_1 H_{\omega,N}=H_{\omega,N}-(1-\tilde{\varphi}_1)H_{\omega,N}=\tilde{G}_{\omega,N}-\tilde{\Psi}_1\tilde{G}_{\omega,N}$$

is the Fourier-transform of $G_\omega=G_{\omega,N}-\Psi_1 * G_{\omega,N}$ and G_ω enjoys the same properties as $G_{\omega,N}$; note that $H_\omega(i\rho)=(i\rho+\omega-A-\widehat{dB}(i\rho))^{-1}$ coincides with $H_{\omega,N}(\rho)$ for $|\rho|\geqq N$ since $(k_N * B_1)\tilde{}\,(\rho)=0$ there. Finally to get from $H_\omega(i\rho)$ to $H(i\rho)$ we consider the equation

$$H(i\rho)=H_\omega(i\rho)+\omega H(i\rho)H_\omega(i\rho).$$

Choosing another cutoff function $\tilde{\varphi}_2\in C^\infty$ with $\tilde{\varphi}_2\equiv 1$ on supp $\tilde{\varphi}_1$, we then have

$$\tilde{\varphi}_1 H=\tilde{\varphi}_1 H_\omega+\omega\tilde{\varphi}_1 H\cdot\tilde{\varphi}_2 H_\omega,$$

where $\tilde{\varphi}_1 H_\omega$ and $\tilde{\varphi}_2 H_\omega$ are Fourier-transforms of functions from $L^1(\mathbb{B}(X))\cap L^\infty(\mathbb{B}(X))$, hence, by Lemma 2, $\tilde{\varphi}_1 H$ is the Fourier-transform of $G_2\in L^1(\mathbb{B}(X))\cap L^\infty(\mathbb{B}(X))$ with the desired properties, since

$$\begin{aligned} I-\omega\tilde{\varphi}_2 H_\omega&=(i\rho+\omega-A-\widehat{dB}-\omega\varphi_2)H_\omega\\ &=[I-\widehat{dB}(i\rho+\omega(1-\tilde{\varphi}_2)-A)^{-1}](i\rho+\omega(1-\tilde{\varphi}_2)-A)H_\omega \end{aligned}$$

is invertible for each $\rho\in\mathbb{R}$.

(iv) Up to this point we have obtained the kernel G_Λ with the desired properties and G_Λ represents the solution operator G by (12); $G_\Lambda\in L^1(\mathbb{B}(X))$ also implies $G\in\mathbb{B}(\Lambda(X))$. To derive the admissibility of $\Lambda(X)$ for (1), by Theorem 1, it thus remains to show that G is also bounded from $\Lambda(X)\cap C_u^1(X)$ to $C_u(Y)$, and for this in turn it suffices to prove that $\tilde{\varphi}_1(\rho)H(i\rho)/i\rho$ is the Fourier-transform of some $V\in L^1(\mathbb{B}(X, Y))$. Since $\tilde{\varphi}_1(\rho)/i\rho$ and $\tilde{\varphi}_1(\rho)H_\omega(i\rho)$ are Fourier-transforms of $L^1(\mathbb{B}(X))$-functions and

$$I-(\omega-A)^{-1}\tilde{\varphi}_2(\rho)\,\widehat{dB}(i\rho)=(\omega-A)^{-1}(\omega-A-\tilde{\varphi}_2(\rho)\,\widehat{dB}(i\rho))$$

is invertible for each $\rho\in\mathbb{R}$, Lemma 2 and the relation

$$\begin{aligned} \tilde{\varphi}_1 H_\omega(i\rho)/i\rho=(\omega-A)^{-1}\tilde{\varphi}_1(\rho)/i\rho-(\omega-A)^{-1}\tilde{\varphi}_1(\rho)H_\omega(i\rho)\\ +(\omega-A)^{-1}\tilde{\varphi}_2(\rho)\,\widehat{dB}(i\rho)\cdot\tilde{\varphi}_1(\rho)H_\omega(i\rho)/i\rho \end{aligned}$$

show that $\tilde{\varphi}_1 H_\omega/i\rho=\tilde{V}_\omega$ for some $V_\omega\in L^1(\mathbb{B}(X, Y))$. Finally, another application of Lemma 2 to

$$\tilde{\varphi}_1 H/i\rho=\tilde{V}_\omega+(\tilde{\varphi}_1 H/i\rho)\cdot\tilde{\varphi}_2\omega H_\omega$$

yields $V\in L^1(\mathbb{B}(X, Y))$ with $\tilde{V}(\rho)=\tilde{\varphi}_1(\rho)H(i\rho)/i\rho$.

The proof is now complete. □

(f) *Proof of Theorem* 3.

(i) The "only if" part is contained in Proposition 1(i). For the "if" part we note first that again dist $(\Lambda, \Lambda_0)\geqq 3\varepsilon>0$ for some $\varepsilon>0$ holds. In fact, the relation $\widehat{B'}(\lambda)=\lambda^{-1}\widehat{(dB')}(\lambda)$ together with the assumptions of Theorem 3 implies

$$|\widehat{B'}(i\rho)y|\leqq C|y|_Y/(1+|\rho|)\quad\text{for all } y\in D(A),\quad \rho\in\mathbb{R}, \tag{29}$$

and therefore

$$i\rho - A - \widehat{B'}(i\rho) = [I + i(\rho - \tau)H(i\tau) + (\widehat{B'}(i\tau) - \widehat{B'}(i\rho))H(i\tau)](i\tau - A - \widehat{B'}(i\tau))$$

shows the invertibility of $i\rho - A - B'(i\rho)$ on dist $(\rho, \Lambda) \leqq 3\varepsilon$. So we may choose a cutoff function $\tilde{\varphi} \in C^\infty$ such that $\tilde{\varphi}(\rho) = 1$ on dist$(\rho_1 \Lambda) \leqq \varepsilon$; w.l.o.g. we may also suppose that $\tilde{\varphi}(\rho) = 0$ for $\rho \in [-\varepsilon_0, \varepsilon_0]$ holds for some $\varepsilon_0 > 0$, otherwise choose another cutoff function $\tilde{\varphi}_1 \in C_0^\infty$ such that $\tilde{\varphi}_1 = 1$ on $[-\varepsilon_0, \varepsilon_0]$, $0 \leqq \tilde{\varphi}_1 \leqq 1$, and consider $(1 - \tilde{\varphi}_1)\tilde{\varphi}$ instead of $\tilde{\varphi}$. The part corresponding to $\tilde{\varphi}_1$

$$G_1(t) = (2\pi)^{-1} \int_{-\infty}^{\infty} \tilde{\varphi}_1(\rho)\tilde{\varphi}(\rho)H(i\rho)\, e^{i\rho t}\, d\rho$$

belongs to $C_b^\infty(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ by Lemma 1.

(ii) Let $\omega \in \mathbb{R}$ be large enough such that $|T(t)| \leqq M e^{(\omega - 1)t}$ for $t > 0$ and some $M > 1$, where $T(t)$ denotes the semigroup generated by A. Then $f_1(t) = h_0(t)\, e^{-\omega t} T(t)x$ as well as $f_2(t) = h_0(t)\, e^{-\omega t} T(t)^* x$ belong to $L^2(X)$, where $h_0(t)$ means Heaviside's function and the star indicates the adjoint. Since X is Hilbert, their Fourier-transforms $\tilde{f}_1(\rho) = (i\rho + \omega - A)^{-1}x$ and $\tilde{f}_2(\rho) = (i\rho + \omega - A)^{-1*}x$ belong to $L^2(X)$. By (29), $K(\rho) = \widehat{B'}(i\rho)H(i\rho)$ is $\mathbb{B}(X)$-continuous and uniformly bounded for dist $(\rho, \Lambda) \leqq 2\varepsilon$, and so the representation

$$H(i\rho)^* x = (I + \omega H(i\rho)^* + K(\rho)^*)(i\rho + \omega - A)^{-1*} x$$

yields $\tilde{\varphi} H^* x \in L^2(X)$, also. Hence there is $g_x^* \in L^2(X)$ such $\widetilde{g_x^*} = \tilde{\varphi} H^* x$. Similarly, the relation

$$A(i\rho + \omega - A)^{-1} x = (i\rho + \omega)(i\rho + \omega - A)^{-1} x - x$$

shows that $\tilde{\varphi}(\rho) A(i\rho + \omega - A)^{-1} x / i\rho$ as well as $\tilde{\varphi}(\rho) A^*(i\rho + \omega - A)^{-1*} x / i\rho$ also belong to $L^2(X)$, and so by means of

$$Hx = (I + \omega H)(i\rho + \omega - A)^{-1} x + H[i\rho \widehat{B'}(\omega - A)^{-1}](\omega - A)(i\rho + \omega - A)^{-1} x / i\rho$$

and

$$(AH)^* x = (I + \omega H^* + K^*) A^*(i\rho + \omega - A)^{-1*} x$$

we see by (29) that $\tilde{\varphi} Hx$ and $\tilde{\varphi}(AH)^* x / i\rho$ belong to $L^2(X)$ and so there are g_x, $f_x^* \in L^2(X)$ such that $\tilde{g}_x = \tilde{\varphi} Hx$ and $\widetilde{f_x^*} = \tilde{\varphi}(AH)^* x / i\rho$ hold in $L^2(X)$. Note that g_x, g_x^*, f_x^* are uniquely determined by $x \in X$ in $L^2(X)$, i.e., almost everywhere.

(iii) Next we show that g_x, g_x^*, f_x^* also belong to $L^1(X)$. Since B' has first moment by assumption, $H(i\rho)$ is differentiable for dist $(\rho, \Lambda) \leqq 2\varepsilon$ and with $\partial = \partial / \partial\rho$ we obtain

$$\partial H = H(i - \partial \widehat{B'})H = H(iH - K_1);$$

hence

$$\partial(\tilde{\varphi} H^* x) = \partial\tilde{\varphi} H^* x + \tilde{\varphi}\, \partial H^* x = \partial\tilde{\varphi} H^* x - (iH^* + K_1^*)\tilde{\varphi} H^* x$$

belongs to $L^2(X)$ since $K_1 = (\partial \widehat{B'})H = -i(B')\hat{}\, H$ is continuous and bounded for dist $(\rho, \Lambda) \leqq 2\varepsilon$. But this shows $t g_x^* \in L^2(X)$ and therefore

$$\int_{-\infty}^{\infty} |g_x^*(t)|\, dt = \int_{|t| \leqq 1} |g_x^*(t)|\, dt + \int_{|t| \geqq 1} t^{-1} |t g_x^*(t)|\, dt$$
$$\leqq |g_x^*|_{L^2} + \sqrt{2}\, |t g_x^*|_{L^2} < \infty,$$

i.e., $g_x^* \in L^1(X)$. Similarly, one also obtains g_x, $f_x^* \in L^1(X)$.

(iv) The mappings $x \to g_x$, $x \to g_x^*$, $x \to f_x^*$ from X to $L^1(X)$ are closed, linear and defined on all X; hence by the closed graph theorem, there is some $C > 0$ such that

$$|g_x|_{L_1} + |g_x^*|_{L^1} + |f_x^*|_{L^1} \leqq C|x| \quad \text{for all } x \in X.$$

To prove the boundedness of the solution operator G from $\Lambda(X)$ to $C_u(X)$ as well as from $\Lambda(X) \cap C_u^1(X)$ to $C_u(Y)$, it therefore suffices to show that the relations

$$\begin{aligned} (x, u(t)) &= \int_{-\infty}^{\infty} (g_x^*(t-\tau), f(\tau))\, d\tau, \qquad t \in \mathbb{R}, \quad x \in X, \\ (x, Au(t)) &= \int_{-\infty}^{\infty} (f_x^*(t-\tau), f'(\tau))\, d\tau, \qquad t \in \mathbb{R}, \quad x \in X \end{aligned} \tag{30}$$

hold for solutions $u \in \Lambda(X) \cap C_u^1(X) \cap C_u(Y)$ of (1) with $f \in \Lambda(X)$, $\sigma(f)$ compact, which exist by Theorem 1.

So let such f be given, choose another cutoff function $\tilde{\varphi}_0 \in C_0^\infty$ such that $\tilde{\varphi}_0(\rho) = 1$ on $\sigma(f)$ and let $G_0 \in C_b(\mathbb{B}(X, Y)) \cap L^1(\mathbb{B}(X, Y))$ from Lemma 1 such that $\tilde{G}_0 = \tilde{\varphi} H \cdot \tilde{\varphi}_0$. Then we have

$$\begin{aligned} (x, u(t)) = (x, G_0 * f(t)) &= \int_{-\infty}^{\infty} (G_0^*(t-\tau)x, f(\tau))\, d\tau \\ &= \int_{-\infty}^{\infty} ((g_x^* * \varphi_0)(t-\tau), f(\tau))\, d\tau = \int_{-\infty}^{\infty} (g_x^*(t-\tau), (\varphi_0 * f)(\tau))\, d\tau \\ &= \int_{-\infty}^{\infty} (g_x^*(t-\tau), f(\tau))\, d\tau \end{aligned}$$

since $\varphi_0 * f = f$ by Proposition 2 and $\widetilde{G_0^* x} = \tilde{\varphi} H^* x \tilde{\varphi}_0 = \widetilde{g_x^*} \cdot \tilde{\varphi}_0 = (g_x^* * \varphi_0)^{\sim}$; hence $G_0^* x = g_x^* * \varphi_0$ by uniqueness of the Fourier-transform. Similarly one obtains the second part of (30).

(v) Suppose Λ_0 is compact, then w.l.o.g. we may choose $\tilde{\varphi}$ such that $\tilde{\varphi}(\rho) = 1$ for $|\rho| \geqq N$, where N is sufficiently large. Since for $x \in D(A)$

$$i\rho H(i\rho)x - x = i\rho(H(i\rho)x - x/i\rho) = H(i\rho)Ax + H(i\rho)\widehat{B'}(i\rho)x$$

belongs to $L^2(X)$ asymptotically, we obtain from

$$\tilde{\varphi}(\rho)H(i\rho)x = \tilde{\varphi}(\rho)(H(i\rho)x - x/i\rho) + (\tilde{\varphi}(\rho) - 1)x/i\rho + x/i\rho$$

that $g_x(t)$ is continuous for $t \neq 0$ and that the jump relation $g_x(0+) - g_x(0-) = x$ is satisfied for each $x \in D(A)$. Also since $\tilde{\varphi} H^* x \in L^2(x)$ and $(i\rho + \omega - A)^{-1}x \in [L^\infty(X)]^{\sim}$ we derive from

$$(x^*, Hx) = (x^*, (i\rho + \omega - A)^{-1}x) + (H^* x, (\omega + \widehat{B'})(i\rho + \omega - A)^{-1}x) \in L^1(X)$$

the boundedness of $(x^*, g_x(t))$ for all $x, x^* \in X$. Hence $g_x(t)$ is bounded in t for each $x \in X$, and therefore continuous for $t \neq 0$ and $g_x(0+) - g_x(0-) = 0$ holds, by density of $D(A)$ in X. This shows that there is an operator-valued kernel $G_\Lambda(\cdot)$ representing $g_x(t)$ by $g_x(t) = G_\Lambda(t)x$. □

REFERENCES

[1] L. AMERIO AND G. PROUSE, *Almost-Periodic Functions and Functional Equations*, Van Nostrand-Reinhold, New York, 1971.

[2] B. D. COLEMAN AND M. E. GURTIN, *Equipresence and constitutive equation for rigid heat conductors*, Z. Angew. Math. Phys., 18 (1967), pp. 199–208.
[3] J. M. CUSHING, *Integrodifferential Equations and Delay Models in Population Dynamics*, Lecture Notes in Biomathematics 20, Springer-Verlag, Berlin, 1977.
[4] C. M. DAFERMOS, *An abstract Volterra equation with applications to linear viscoelasticity*, J. Differential Equations, 7 (1970), pp. 554–569.
[5] A. FRIEDMAN AND M. SHINBROT, *Volterra integral equations in Banach space*, Trans. Amer. Math. Soc., 126 (1967), pp. 131–179.
[6] G. GREINER, J. VOIGT AND M. WOLFF, *On the spectral bound of the generator of a semigroup of positive operators*, J. Operator Theory, 5 (1981), pp. 245–256.
[7] R. GRIMMER AND J. PRÜSS, *On linear Volterra equations in Banach spaces*, Comput. Math. Appl., 11 (1985), pp. 189–205.
[8] G. GRIPENBERG, *Asymptotic behaviour of resolvents of abstract Volterra equations*, to appear.
[9] G. HAGEDORN, *Asymptotic completeness for the impact parameter approximation to three particle scattering*, Ann. Inst. H. Poincaré Sect. A, 36 (1982), pp. 19–40.
[10] E. HILLE AND R. S. PHILLIPS, *Semigroups and Functional Analysis*, American Mathematical Society Colloq. Publ. 31, Providence, RI, 1957.
[11] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, 2nd corrected ed., Dover Books on Advanced Mathematics, New York, 1976.
[12] R. K. MILLER AND R. L. WHEELER, *Asymptotic behavior for a linear Volterra integral equation in Hilbert space*, J. Differential Equations, 23 (1977), pp. 270–284.
[13] J. PRÜSS, *Linear Volterra Gleichungen un Banachräumen*, Habilitationsschrift, Universität Paderborn, Paderborn, West Germany, 1984.
[14] ———, *On Volterra equations of parabolic type in Banach spaces*, Trans. Amer. Math. Soc., to appear.
[15] H. TANABE, *Equations of Evolution*, Monographs and Studies in Mathematics 6, Pitman, London, 1979.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

011

ON LIMIT STATES OF A LINEARIZED BOLTZMANN EQUATION*

MILAN MIKLAVČIČ†

Abstract. With methods of mean ergodic theory a very simple criterion for existence of limit states of linearized Boltzmann equation is proven.

Key words. linearized Boltzmann equation, ergodic theory

AMS(MOS) subject classifications. 45A05, 47A35

A linearized Boltzmann equation can under some conditions ([4], see also [1], [8]) be written in the form

$$\frac{du}{dt} + u = Bu$$

where B is typically a Markov operator on some L^1. The solution of this equation is $e^{-t}e^{Bt}u_0$ and one would like to know how the solution behaves as $t \to \infty$. In this paper this equation is considered in an arbitrary Banach space X and B is assumed to be a bounded linear operator on X such that $\sup_{n\geqq 0}\|B^n\| < \infty$. Let $N = \{x \in X \mid Bx = x\}$ and $R = \{x \in X \mid x = y - By \text{ for some } y \in X\}$.

THEOREM 1. *If $x \in X$, $x_0 \in X$, then the following statements are equivalent*:

(a) *There exist integers $n_1 < n_2 < \cdots$ such that $\lim_{i\to\infty} y((1/n_i)\sum_{k=0}^{n_i-1} B^k x) = y(x_0)$ for all $y \in X^*$.*

(b) *There exist $t_1, t_2, \cdots$ in $(0, \infty)$ such that $\lim_{i\to\infty} t_i = \infty$ and $\lim_{i\to\infty} y(e^{-t_i}e^{Bt_i}x) = y(x_0)$ for all $y \in X^*$.*

(c) $x_0 \in N$ *and* $x - x_0 \in \bar{R}$.

(d) $\lim_{n\to\infty} \|(1/n)\sum_{k=0}^{n-1} B^k x - x_0\| = 0$.

(e) $\lim_{t\to\infty} \|e^{-t}e^{Bt}x - x_0\| = 0$.

This theorem implies that all standard mean ergodic theorems (e.g. (2)) are applicable in the study of limits of $e^{-t}e^{Bt}x$! A simple and quite powerful criterion for existence of the limit is given by the following.

THEOREM 2. *If $x \in X$ and if the set $\{B^n x \mid n \geqq 0\}$ is weakly sequentially compact, then there exists $x_0 \in N$ such that*

$$\lim_{t\to\infty} \|e^{-t}e^{Bt}x - x_0\| = 0.$$

Proof. Let $C = \{B^n x \mid n \geqq 0\}$ and let C_1 be the convex hull of C. Since $(1/n)\sum_{k=0}^{n-1} B^k x \in C_1$ for $n \geqq 1$ and since C_1 is weakly sequentially compact [2, Krein-Šmulian Theorem] there exists $x_0 \in X$ such that part (a) of Theorem 1 is satisfied.

There are many ways of showing that the set $\{B^n x \mid n \geqq 0\}$ is weakly sequentially compact [2]. In [3], [5] conditions like strong (weak) constrictiveness (and some other conditions) are required; the following observation shows that these conditions are much more restrictive than those of Theorem 2. If $x \in X$, then the set $\{B^n x \mid n \geqq 0\}$ is weakly sequentially compact if and only if there exists a weakly compact set F (for this x!) such that $\lim_{n\to\infty} \operatorname{dist}(B^n x, F) = 0$.

If B is positive quasi-compact operator then even convergence rates can be estimated.

* Received by the editors March 31, 1986; accepted for publication February 9, 1987.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824.

THEOREM 3. *Suppose that X is real Banach space and that X^+ is a closed subset of X with the following properties*:

(1) *If $x \in X^+$, $y \in X^+$, $\alpha \in [0, \infty)$ then $x+y \in X^+$ and $\alpha x \in X^+$;*

(2) *There exists $M_0 \in (0, \infty)$ such that for each $x \in X$ there exist $x_+ \in X^+$ and $x_- \in X^+$ which satisfy*

$$x = x_+ - x_-, \quad \|x_+\| \leqq M_0\|x\|, \quad \|x_-\| \leqq M_0\|x\|$$

and if $x = y_+ - y_-$ for some $y_+ \in X^+$, $y_- \in X^+$, then $y_+ - x_+ \in X^+$.

(3) *If $x \in X^+$, $y \in X^+$, then $\|x\| \leqq \|x+y\|$.*

Suppose also that T is bounded linear operator on X such that

(4) *$TX^+ \subset X^+$.*

(5) *$\lim_{n\to\infty} (1/n) y(T^n x) = 0$ for all $x \in X$ and all $y \in X^*$.*

(6) *$\|T^m - K\| < 1$ for some integer m and some compact linear operator K.*

Then there exist $a \in (0, \infty)$, $b \in (0, \infty)$ such that for every $x \in X$ there exists $x_0 \in X$ for which

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} T^k x - x_0 \right\| \leqq \frac{b}{n}\|x\|, \qquad \|e^{-t} e^{Tt} x - x_0\| \leqq b\|x\| \, e^{-at}$$

whenever $n \geqq 1$, $t > 0$.

Proof of Theorem 3 can be found in [7] and follows from the fact that $T^n x$ is actually asymptotically periodic for every $x \in X$ (see also [6]). Operators considered in [5] in connection with the Boltzmann equation satisfy above assumptions.

Proof of Theorem 1. Equivalence of statements (a), (c), (d) is well known [2]. It is obvious that (e) implies (b) and (c); it is enough to prove that (b) implies (c) and that (c) implies (e).

Let $M = \sup_{n\geqq 0} \|B^n\|$, $A = I - B$ and

$$F(t) = e^{-t} + e^{-t} \sum_{n=0}^{\infty} \frac{t^n}{n!} \left| \frac{t}{n+1} - 1 \right|.$$

Identities

$$e^{-At} = e^{-t} e^{Bt} = e^{-t} \sum_{n=0}^{\infty} \frac{t^n}{n!} B^n,$$

$$Ae^{-At} = e^{-t} + e^{-t} \sum_{n=0}^{\infty} \frac{t^n}{n!} \left(\frac{t}{n+1} - 1 \right) B^{n+1}$$

imply that for $t > 0$

$$\|e^{-At}\| \leqq M, \qquad \|Ae^{-At}\| \leqq MF(t).$$

A nasty but straightforward exercise gives that $\lim_{t\to\infty} F(t) = 0$.

Assume (b). Then for all $y \in X^*$

$$0 = \lim_{i\to\infty} y(Ae^{-At_i} x) = \lim_{i\to\infty} (A^* y)(e^{-At_i} x) = (A^* y)(x_0) = y(Ax_0);$$

hence, $Ax_0 = 0$. If $x - x_0 \notin \bar{R}$, then there exists $y_0 \in X^*$ such that $A^* y_0 = 0$ and $y_0(x - x_0) = 1$, which leads to the contradiction

$$1 = y_0(x - x_0) = (e^{-A^* t_i} y_0)(x - x_0) = y_0(e^{-At_i} x - x_0) = 0$$

and therefore $x - x_0 \in \bar{R}$ and (c) is true.

Assume (c). Pick $\varepsilon > 0$ and let $z \in X$ be such that $\|x - x_0 - Az\| < \varepsilon/(1+M)$; hence, for $t > 0$

$$\begin{aligned}\|e^{-At}x - x_0\| &\leqq \|e^{-At}(x - x_0 - Az)\| + \|Ae^{-At}z\| \\ &\leqq M\varepsilon/(1+M) + MF(t)\|z\|\end{aligned}$$

and this implies (e).

Acknowledgment. It is my pleasure to thank T. Y. Li for many illuminating discussions.

REFERENCES

[1] M. F. BARNSLEY AND G. TURCHETTI, *On the Abel transformation and the nonlinear Boltzmann equation*, Phys. Lett. A, 72 (1979), pp. 417-419.

[2] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* 1, Wiley-Interscience, New York, 1967.

[3] J. KOMORNIK, *Asymptotic periodicity of the iterates of weakly constrictive Markov operators*, Tôhoku Math. J., to appear.

[4] A. LASOTA, *Statistical stability of deterministic systems*, Lecture Notes in Mathematics, 1017, Springer-Verlag, New York, 1983, pp. 386-419.

[5] A. LASOTA, T. Y. LI AND J. A. YORKE, *Asymptotic periodicity of the iterates of Markov operators*, AMS Transactions, 286 (1984), pp. 751-764.

[6] M. LIN, *Quasi-compactness and uniform ergodicity of positive operators*, Israel J. Math., 29 (1978), pp. 309-311.

[7] M. MIKLAVČIČ, *Asymptotic periodicity of the iterates of positivity preserving operators*, AMS Transactions, submitted.

[8] J. A. TJON AND T. T. WU, *Numerical aspects of the approach to a Maxwellian-distribution*, Phys. Rev. A., 19 (1979), pp. 883-888.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

012

EXISTENCE, UNIQUENESS AND REGULARITY OF A TIME-PERIODIC PROBABILITY DENSITY DISTRIBUTION ARISING IN A SEDIMENTATION-DIFFUSION PROBLEM*

LUDWIG C. NITSCHE†‡, JOHANNES M. NITSCHE†‡ AND HOWARD BRENNER†§

Abstract. An analysis is presented of the one-dimensional convective-diffusive equation governing the temporal evolution and spatial distribution of the probability density $P(x, t)$ describing the sedimentation and diffusion of a nonneutrally buoyant Brownian particle in a vertical fluid-filled cylinder that is flipped over instantaneously at regular intervals. A time-periodic solution is sought by requiring that the initial spatial distribution recur after one complete period of the flipping motion. The periodicity condition is formulated both as an infinite matrix equation for the eigenfunction expansion coefficients of a possible recurring initial distribution, and as an integral equation of the second kind for the distribution itself. The kernel (infinite matrix) is shown to be square integrable (square summable), so that Fredholm theory applies. There exists a unique time-periodic solution whenever unity is not an eigenvalue of the integral (infinite matrix) operator. Regions in parameter space are identified in which existence and uniqueness are assured. It is shown that time-periodic solutions are analytic functions of the position coordinate at each time. A symmetry property of unique time-periodic solutions is established by deriving a simpler infinite matrix equation. Numerical examples show the appearance of recurring initial distributions.

Key words. sedimentation-diffusion, time-periodic solution, Fredholm theory, existence, uniqueness, regularity

AMS(MOS) subject classifications. primary 35B10, 35K99, 80A20; secondary 35R05, 45B05

1. Introduction.

Motivation. (Small) nonneutrally buoyant particles sedimenting under the influence of gravity in an otherwise quiescent viscous suspension ultimately settle to the bottom of the container. This phenomenon unfortunately eliminates the possibility of a leisurely experimental study of the physicochemical, colloidal or biophysical properties of individual particles in their supernatant fluid environment. While the suspension can, of course, be stirred to levitate the particles, the relatively large shear fields engendered by the stirring may fundamentally alter the intrinsic particle property being studied, e.g., the configuration of the body in the case of flexible or deformable particles. Even when this is not the case, the introduction of largely unknown velocity fields by the agitator may be expected to complicate the interpretation of any experimental results.

With these difficulties in mind, a novel scheme for achieving (time-average) nonsedimenting states for small nonneutrally buoyant particles has recently been proposed by Dill and Brenner [5], and further elaborated by Nadim, Cox and Brenner [15] and Davis and Brenner [4]. This involves placing the suspension into a *slowly* rotating horizontal circular cylinder, so that the suspension as a whole undergoes a rigid-body rotation upon which is superposed the Stokes settling velocity of each particle with respect to the fluid in its instantaneous locale. (*Slow* rotation assures the absence of centrifugal and Coriolis effects, whose presence would negate the theoretical basis of the proposed scheme.) This combined translational/rotational particle motion

* Received by the editors August 4, 1986; accepted for publication February 9, 1987. This paper was presented at the SIAM 1986 National Meeting, July 21-25, 1986, Boston, Massachusetts.

† Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ The work of this author was supported by a National Science Foundation Graduate Fellowship.

§ The work of this author was supported by the Microgravity Sciences and Applications Program of the National Aeronautics and Space Administration through grant NSG-7645 administered by the Materials Processing Center of the Massachusetts Institute of Technology.

relative to a space-fixed observer results in a closed, time-periodic particle trajectory, whose *net* motion over one period is identically zero [5]. (This same conclusion obtains [15] even when the sedimenting particle is small enough to undergo appreciable translational Brownian motion, whereupon its trajectory is now stochastic rather than deterministic—the former circumstance being the case of interest to us in the subsequent analysis.) Thus, despite the particle's continuous *relative* settling, it does not undergo any *net* settling relative to a space-fixed observer.

The underlying mechanism behind this class of "antisedimentation" devices is most readily visualized by imagining oneself to be a body-fixed observer, fixed in a reference frame that rotates with the fluid-filled cylinder. From this vantage point the otherwise constant, space-fixed gravity force vector acting on each particle is observed to be a time-periodic vector (of constant magnitude), whose average value over one period is the zero vector. In effect, it is this zero-valued mean which is the source of the state of levitation. It effectively nullifies the action of gravity [16].

In a specific context the class of such time-periodic, zero-mean sedimentation phenomena may be simulated, both physically and mathematically, by allowing the suspension to settle in a (nonrotating) vertical cylinder that is flipped over at periodic intervals. (Any sediment near to the bottom of the cylinder then suddenly finds itself near to the top, whereupon it must again settle through the entire length of the cylinder before reaching the bottom, et cycl.) An observer fixed in the cylinder walls will evidently observe the gravity-force vector to be time-periodic with zero mean. The physicomathematical question to be addressed here is whether or not this zero-mean, time-periodic external force will give rise to a zero-mean, time-periodic particle displacement—and hence a levitated particle state. More explicitly, we address this question in the specific circumstances of interest to us in applications, namely that the sedimenting particles be small enough to also undergo appreciable diffusion. Because of this small particle size, both particle and fluid inertial effects can be ignored, and the instantaneous particle settling velocity assigned its quasistatic Stokes law value.

Formulation of the physical problem. Consider the sedimentation and diffusion of a nonneutrally buoyant Brownian particle (or, equivalently, a dilute sedimenting suspension of identical nonhydrodynamically-interacting such particles) in a vertical fluid-filled cylinder of finite length l that is flipped over instantaneously at regular intervals Θ. In a reference frame fixed to the container the particle's settling velocity vector alternately points up and down. The analysis is simplified by considering a one-dimensional description, in which the only spatial coordinate is distance x parallel to the cylinder axis. Then the single-body probability density $P(x, t)$ for the particle position x at time t is governed by a standard one-dimensional convective-diffusion equation, together with no-flux boundary conditions at the ends $x = 0, l$ of the container, and an arbitrarily prescribed initial spatial distribution $f(x)$ at $t = 0$. These equations may be cast in the following dimensionless form (with x, t, P and f henceforth nondimensional):

$$\left.\begin{aligned}
&\frac{\partial P}{\partial t}+u(t)\frac{\partial P}{\partial x}=\frac{1}{b}\frac{\partial^2 P}{\partial x^2}, && 0<x<1, \quad t>0, && (1.1)\\
&\frac{\partial P}{\partial x}(0,t)-bu(t)P(0,t)=0, && t>0, && (1.2)\\
&\frac{\partial P}{\partial x}(1,t)-bu(t)P(1,t)=0, && t>0, && (1.3)\\
&P(x,0)=f(x), && 0<x<1, && (1.4)
\end{aligned}\right\}\ \mathscr{P}$$

where

$$u(t)=\begin{cases}1, & 0<t<T,\\ -1, & T<t<2T,\end{cases} \tag{1.5a}$$

$$u(t+2T)=u(t) \tag{1.5b}$$

and

$$\int_0^1 f(x)\,dx=1 \qquad \text{(normalization condition).} \tag{1.6}$$

The two positive nondimensional parameters appearing in this formulation are the Péclet number b and dimensionless half-period T, respectively defined as

$$b=lc/D, \tag{1.7}$$

$$T=\Theta c/l, \tag{1.8}$$

with l the container length, $c>0$ the Stokes' law settling speed of the particle, $D>0$ its Brownian diffusivity through the fluid, and Θ the time interval between successive overturnings.

Considering the time periodicity of the dimensionless particle settling velocity $u(t)$, it is natural to inquire whether or not problem $\mathcal{P}$ possesses a (square integrable) time-periodic solution; equivalently, does there exist an "initial" distribution $P(x,0)=f(x)$ which will recur at time $t=2T$? It is the purpose of this paper to prove the existence and uniqueness of time-periodic solutions for various ranges of the parameters b and T. Moreover, we address the question of regularity by showing that time-periodic solutions are analytic functions of x at each time t. We also establish a time-shifted symmetry property of unique time-periodic solutions. Illustrations are appended which show numerically-generated recurring initial distributions.

Although the literature on the general subject of time-periodic solutions of parabolic differential equations is considerable, the present problem seems not to have been treated before. Different types of boundary conditions have been considered; Dirichlet conditions seem to predominate. There have appeared many analyses of problems which, in contrast with the present problem $\mathcal{P}$, involve time periodicity only in the differential operator or only in the boundary conditions, but not in both (e.g. [11]). Extensive bibliographies are available in Cannon [2] and Vejvoda [21]. Differential equations which could reduce to an equation similar to (1.1), together with time-periodic boundary conditions involving the unknown function and a spatial derivative thereof, are considered by Farlow [6], [7], Heuss [9], Knolle [12], Mal'tsev [13], Prodi [17] and Šmulev [19], [20]. The principal differences between these analyses and the present one are that: (i) In contrast to (1.2) and (1.3), the boundary conditions in the aforementioned papers are of the conventional type, which require P and its normal (or oblique) derivative with a consistent orientation to have the same sign relationship at all points of the boundary; (ii) the time-periodic functions considered in the cited references generally must satisfy certain regularity conditions, whereas our time-periodic velocity $u(t)$ is discontinuous.

2. Equations for a time-periodic solution. A time-periodic solution of problem $\mathcal{P}$ is sought by requiring that the initial distribution $f(x)=P(x,0)$ recur after the system evolves through one complete period of the flipping motion; explicitly, $f(x)$ should satisfy

$$f(x)=P(x,0)=P(x,2T), \qquad 0\leqq x\leqq 1. \tag{2.1}$$

Since $u(t)$ is time independent during each half period, problem $\mathcal{P}$ may be subdivided into two constant-coefficient subproblems, $\mathcal{P}^+$ and $\mathcal{P}^-$. Thus, the probability density $P(x, t)$ is governed by the set of equations

$$\left.\begin{aligned}
&\frac{\partial P}{\partial t}+\frac{\partial P}{\partial x}=\frac{1}{b}\frac{\partial^2 P}{\partial x^2}, && 0<x<1, \quad 0<t<T, && (2.2)\\
&\frac{\partial P}{\partial x}(0, t)-bP(0, t)=0, && 0<t<T, && (2.3)\\
&\frac{\partial P}{\partial x}(1, t)-bP(1, t)=0, && 0<t<T, && (2.4)\\
&P(x, 0)=f(x), && 0<x<1, && (2.5)
\end{aligned}\right\} \mathcal{P}^+$$

where

$$\int_0^1 f(x)\,dx=1, \tag{2.6}$$

together with

$$\left.\begin{aligned}
&\frac{\partial P}{\partial t}-\frac{\partial P}{\partial x}=\frac{1}{b}\frac{\partial^2 P}{\partial x^2}, && 0<x<1, \quad T<t<2T, && (2.7)\\
&\frac{\partial P}{\partial x}(0, t)+bP(0, t)=0, && T<t<2T, && (2.8)\\
&\frac{\partial P}{\partial x}(1, t)+bP(1, t)=0, && T<t<2T. && (2.9)
\end{aligned}\right\} \mathcal{P}^-$$

Here, the initial distribution $P(x, T)$ for problem $\mathcal{P}^-$ is furnished by the solution of problem $\mathcal{P}^+$ at time $t=T$. It is redundant to demand satisfaction of the unit normalization condition (2.6) by $P(x, T)$ inasmuch as problem $\mathcal{P}^+$ conserves probability density (as do problems $\mathcal{P}^-$ and $\mathcal{P}$, also).

Problems $\mathcal{P}^+$ and $\mathcal{P}^-$ can be solved by separation of variables. Though both lead to the same discrete eigenvalue spectrum,

$$\lambda_0=0, \qquad \lambda_n=\frac{b}{4}+\frac{n^2\pi^2}{b} \quad \text{for } n=1, 2, 3, \cdots, \tag{2.10}$$

the sequence of eigenfunctions is, nevertheless, different for each. With $^+$ and $^-$ superscripts respectively distinguishing problems $\mathcal{P}^+$ and $\mathcal{P}^-$, the eigenfunctions $\phi_n^{\pm}(x)$ are given by

$$\phi_0^{\pm}(x)=\left(\frac{\pm b}{e^{\pm b}-1}\right)^{1/2} e^{\pm bx}, \tag{2.11a}$$

$$\phi_n^{\pm}(x)=\left(\frac{2}{b\lambda_n}\right)^{1/2}\left[n\pi\cos(n\pi x)\pm\frac{b}{2}\sin(n\pi x)\right] e^{\pm bx/2}, \qquad n\geqq 1, \tag{2.11b}$$

and satisfy the orthonormality condition

$$\int_0^1 \phi_m^{\pm}(x)\phi_n^{\pm}(x)\, e^{\mp bx}\,dx=\delta_{mn}, \qquad m, n\geqq 0. \tag{2.12}$$

Both sets of eigenfunctions are complete. A direct computation shows that the two kinds of eigenfunctions are related by the equation

$$\phi_n^-(x) = (-1)^n e^{-b/2}\phi_n^+(1-x) \tag{2.13}$$

for all $n \geqq 0$.

Subsequent analysis necessitates determining the expansion of each function $\phi_n^-(x)$ in terms of the set of functions $\{\phi_n^+(x)\}$ and conversely:

$$\phi_n^-(x) = \sum_{m=0}^{\infty} A_{mn}\phi_m^+(x), \tag{2.14}$$

$$\phi_n^+(x) = \sum_{m=0}^{\infty} B_{mn}\phi_m^-(x) \tag{2.15}$$

($n \geqq 0$), where

$$A_{mn} = \int_0^1 \phi_n^-(x)\phi_m^+(x)\, e^{-bx}\, dx, \tag{2.16}$$

$$B_{mn} = \int_0^1 \phi_n^+(x)\phi_m^-(x)\, e^{bx}\, dx \tag{2.17}$$

($m, n \geqq 0$). By making the substitution $x = 1 - y$ in (2.17), and utilizing (2.13) and (2.16), it can be shown that

$$B_{mn} = (-1)^{m+n} e^b A_{mn}. \tag{2.18}$$

Straightforward, albeit lengthy, computations yield:

$$A_{mn} = \frac{4\pi^2 b[1-(-1)^{m+n} e^{-b}](\lambda_m/\lambda_n)^{1/2} mn}{[b^2+(m+n)^2\pi^2][b^2+(m-n)^2\pi^2]}, \qquad m, n \geqq 1, \tag{2.19a}$$

$$A_{m0} = \left[\frac{2}{\lambda_m(1-e^{-b})}\right]^{1/2} [1-(-1)^m e^{-3b/2}] \frac{8\pi mb}{9b^2+4m^2\pi^2}, \qquad m \geqq 1, \tag{2.19b}$$

$$A_{0n} = e^{-b/2}\delta_{0n}, \qquad n \geqq 0. \tag{2.19c}$$

Explicit expressions for the B_{mn} can be obtained from (2.19) upon using the relation (2.18).

The preceding development permits derivation of the pertinent equations that a recurring initial distribution must necessarily satisfy. Start with some arbitrary initial distribution,

$$P(x, 0) = f(x) = \sum_{m=0}^{\infty} a_m\phi_m^+(x), \qquad 0 \leqq x \leqq 1, \tag{2.20}$$

where, for all $m \geqq 0$,

$$a_m = \int_0^1 f(x)\phi_m^+(x)\, e^{-bx}\, dx. \tag{2.21}$$

Correspondingly, the solution of problem $\mathcal{P}^+$ is given by the series

$$P(x, t) = \sum_{m=0}^{\infty} a_m\, e^{-\lambda_m t}\phi_m^+(x), \qquad 0 \leqq t \leqq T. \tag{2.22}$$

In particular,

$$P(x, T) = \sum_{m=0}^{\infty} a_m\, e^{-\lambda_m T}\phi_m^+(x). \tag{2.23}$$

Since $P(x, T)$ serves as the initial distribution for problem $\mathscr{P}^-$ it is convenient to expand it in terms of the $\phi_k^-(x)$ as

$$P(x, T) = \sum_{k=0}^{\infty} b_k \phi_k^-(x). \tag{2.24}$$

Here, for all $k \geqq 0$,

$$b_k = \int_0^1 P(x, T)\phi_k^-(x)\, e^{bx}\, dx. \tag{2.25}$$

Problem $\mathscr{P}^-$ possesses the solution

$$P(x, t) = \sum_{k=0}^{\infty} b_k\, e^{-\lambda_k(t-T)}\phi_k^-(x), \qquad T \leqq t \leqq 2T. \tag{2.26}$$

The function $f(x)$ is a recurring initial distribution if, and only if,

$$f(x) \equiv P(x, 2T) = \sum_{k=0}^{\infty} b_k\, e^{-\lambda_k T}\phi_k^-(x). \tag{2.27}$$

Introduction of (2.27) into (2.21) leads to the following expression for each coefficient a_m in terms of the coefficients b_k:

$$a_m = \sum_{k=0}^{\infty} A_{mk}\, e^{-\lambda_k T} b_k, \qquad m \geqq 0. \tag{2.28}$$

A similar combination of (2.23) and (2.25) gives

$$b_k = \sum_{n=0}^{\infty} B_{kn}\, e^{-\lambda_n T} a_n, \qquad k \geqq 0. \tag{2.29}$$

Equations (2.28) and (2.29) coalesce into the following infinite system of linear equations for the eigenfunction expansion coefficients a_m:

$$a_m = \sum_{n=0}^{\infty} \left[\sum_{k=0}^{\infty} A_{mk} B_{kn}\, e^{-(\lambda_k + \lambda_n)T} \right] a_n, \qquad m \geqq 0. \tag{2.30}$$

For $m = 0$, (2.30) reduces to the trivial identity $a_0 = a_0$, owing to (2.19c) and the corresponding equation for the B_{0n}. The system (2.30) therefore needs to be considered only for $m \geqq 1$; it may be rewritten as

$$a_m = \sum_{n=1}^{\infty} C_{mn} a_n + R_m a_0, \qquad m \geqq 1, \tag{2.31}$$

where

$$C_{mn} = \sum_{k=1}^{\infty} A_{mk} B_{kn}\, e^{-(\lambda_k + \lambda_n)T}, \tag{2.32}$$

$$R_m = A_{m0}\, e^{b/2} + \sum_{k=1}^{\infty} A_{mk} B_{k0}\, e^{-\lambda_k T} \qquad (m, n \geqq 1). \tag{2.33}$$

The coefficient a_0 of the expansion (2.20), which appears on the right-hand side of (2.31), must still be specified. Direct calculation gives

$$\int_0^1 \phi_m^+(x)\, dx = \left(\frac{e^b - 1}{b} \right)^{1/2} \delta_{0m}, \tag{2.34}$$

whence the normalization condition (2.6) involves only a_0, thereby yielding

$$a_0 = \left(\frac{b}{e^b - 1}\right)^{1/2}. \tag{2.35}$$

Equation (2.31) constitutes an infinite system of linear equations for the eigenfunction expansion coefficients a_n $(n \geqq 1)$ of the possible recurring initial distribution(s). The question of existence and uniqueness of a square integrable time-periodic solution of problem $\mathscr{P}$ is equivalent to the question of whether or not the infinite linear system (2.31) possesses a unique square summable solution.

There also exists an equivalent integral equation of the second kind for the recurring initial distributions themselves. This may be derived by first reformulating the solutions of problems $\mathscr{P}^+$ and $\mathscr{P}^-$ in terms of their respective Green's functions $G^+(x, t; \xi)$ and $G^-(x, t; \xi)$. In particular,

$$P(x, T) = \int_0^1 G^+(x, T; \xi) f(\xi)\, d\xi. \tag{2.36}$$

With this as the initial distribution for problem $\mathscr{P}^-$, we obtain

$$\begin{aligned} P(x, 2T) &= \int_0^1 G^-(x, 2T; \eta) P(\eta, T)\, d\eta \\ &= \int_0^1 G^-(x, 2T; \eta) \left[\int_0^1 G^+(\eta, T; \xi) f(\xi)\, d\xi \right] d\eta. \end{aligned} \tag{2.37}$$

Accordingly, the periodicity condition (2.1) may be expressed in the form

$$f(x) = P(x, 2T) = \int_0^1 H(x; \xi) f(\xi)\, d\xi, \tag{2.38}$$

where

$$H(x; \xi) = \int_0^1 G^-(x, 2T; \eta) G^+(\eta, T; \xi)\, d\eta. \tag{2.39}$$

Series representations for the Green's functions may be derived from (2.21), (2.22), (2.25) and (2.26). Using these, we ultimately obtain

$$H(x; \xi) = K(x; \xi) + Q(x), \tag{2.40}$$

where

$$K(x; \xi) = e^{-b\xi} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} B_{mn}\, e^{-(\lambda_m + \lambda_n)T} \phi_m^-(x) \phi_n^+(\xi), \tag{2.41}$$

$$Q(x) = \left(\frac{b}{1 - e^{-b}}\right) e^{-bx} + \left(\frac{b}{e^b - 1}\right)^{1/2} \sum_{m=1}^{\infty} B_{m0}\, e^{-\lambda_m T} \phi_m^-(x). \tag{2.42}$$

Substitution of (2.40) into (2.38) together with use of the normalization condition (2.6) leads to an integral equation of the second kind,

$$f(x) = \int_0^1 K(x; \xi) f(\xi)\, d\xi + Q(x), \tag{2.43}$$

for $f(x)$. The last step, whereby (2.38) is rewritten as the inhomogeneous equation (2.43), is analogous to our previous reduction of the homogeneous system (2.30) to the inhomogeneous system (2.31).

We remark that substitution of the series representations (2.20), (2.41) and (2.42) into (2.43) leads to the infinite system (2.31).

3. Existence and uniqueness of time-periodic solutions. Possible conclusions that can be drawn from the integral equation (2.43) necessarily depend upon the properties of the kernel $K(x;\xi)$ as well as those of the inhomogeneous term $Q(x)$. As shown a fortiori by our proof of Lemma 3.2 below, $K(x;\xi)$ is square integrable. Similar arguments establish that $Q(x)$ is also square integrable. It follows from these facts that the standard Fredholm theorems (cf. Mikhlin [14, pp. 1-3, 64-68], Courant and Hilbert [3, pp. 112-122, 152-153]) apply to (2.43).[1] A parallel theory (cf. Hilbert [10, pp. VI-VII], Courant and Hilbert [3, pp. 160-161]) applies to (2.31) because the infinite matrix $C=[C_{mn}]$ and inhomogeneous term $R=[R_m]$ are square summable. Thus, we conclude that there exists a unique (square integrable) recurring initial distribution (equivalently, a unique time-periodic solution) if unity is not an eigenvalue of the integral operator $\mathscr{L}_K=\int_0^1 d\xi\, K(x;\xi)$ or of the infinite matrix $C=[C_{mn}]$.[2] Accordingly, we are motivated to identify regions of the (b, T) parameter plane in which it is impossible for unity to be an eigenvalue. Two independent arguments are presented for this purpose.

LEMMA 3.1. *When*

$$T>2b^2/(b^2+4\pi^2), \tag{3.1}$$

problem $\mathscr{P}$ possesses a unique time-periodic solution.

Proof.[3] Suppose $a_0=0$. Equation (2.29) with $k=0$, in conjunction with (2.19c) and (2.18), then gives $b_0=0$. From (2.24), (2.12) and (2.23) there follows the Parseval identity

$$\sum_{n=1}^{\infty} b_n^2=\int_0^1 e^{bx}\left[\sum_{n=1}^{\infty} a_n e^{-\lambda_n T}\phi_n^+(x)\right]^2 dx, \tag{3.2}$$

in which the summations start from $n=1$ because $a_0=b_0=0$. Straightforward manipulation of the right-hand side, utilizing the bound $e^{bx}\leqq e^{2b}e^{-bx}$ $(0\leqq x\leqq 1)$ together with (2.12), gives

$$\sum_{n=1}^{\infty} b_n^2\leqq e^{2b}\sum_{n=1}^{\infty} a_n^2 e^{-2\lambda_n T}\leqq e^{2(b-\lambda_1 T)}\sum_{n=1}^{\infty} a_n^2. \tag{3.3}$$

A similar argument starting from (2.20), (2.12) and (2.27) shows that

$$\sum_{n=1}^{\infty} a_n^2\leqq e^{-2\lambda_1 T}\sum_{n=1}^{\infty} b_n^2. \tag{3.4}$$

Relations (3.3) and (3.4) can be combined to give

$$\sum_{n=1}^{\infty} a_n^2\leqq e^{2(b-2\lambda_1 T)}\sum_{n=1}^{\infty} a_n^2. \tag{3.5}$$

---

[1] Actually, square integrability of the kernel $K(x;\xi)$ is a sufficient condition for complete continuity of the integral operator. In turn, this property is all that is required (insofar as the integral operator is concerned) to insure the validity of the Fredholm theorems (see Hellinger and Toeplitz [8, p. 1399ff], Riesz and Sz.-Nagy [18, pp. 177-190]).

[2] An eigenvalue μ of the (integral/infinite matrix) operator $\mathscr{L}$ is defined by the equation $y=\mu\mathscr{L}y$, in which y is the corresponding eigenfunction/eigenvector.

[3] This argument was developed in an article by L.C. Nitsche entitled *Settling of a Brownian particle in a container which is flipped over at regular intervals.* (Term paper for the course "Macrotransport Processes," Catalog No. 10.54, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 17, 1985.)

Suppose $T > b/(2\lambda_1) = 2b^2/(b^2+4\pi^2)$. Inequality (3.5) is then satisfied iff $a_n = 0$ for all $n \geqq 1$. Thus, $a_0 = 0 \Rightarrow a_n = 0$ for all $n \geqq 1$. This means that the homogeneous system corresponding to (2.31) has only the trivial solution, i.e., unity is not an eigenvalue of C. It follows as an immediate consequence that problem $\mathscr{P}$ possesses a unique time-periodic solution. Q.E.D.

Another development is based on the fact that the magnitude of an eigenvalue cannot be less than the reciprocal of any norm of the operator; thus, if a norm of the operator is less than unity, no eigenvalue can possibly equal unity. We use this fact in conjunction with a bound on the L^2 norm of $\mathscr{L}_K$ to prove the following lemma.

LEMMA 3.2. *If*

$$\Gamma(b, T) \overset{\text{def}}{=} (e^b-1)^4(\pi bT)^{-2}\, e^{-b(2+T)} \leqq 1, \tag{3.6}$$

then problem $\mathscr{P}$ possesses a unique time-periodic solution.

Proof.

$$\begin{aligned}
\|\mathscr{L}_K\|_2^2 &= \int_0^1 \int_0^1 [K(x;\xi)]^2\, d\xi\, dx \\
&= \int_0^1 \int_0^1 \left[\sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} B_{kl}B_{mn}\, e^{-(\lambda_k+\lambda_l+\lambda_m+\lambda_n)T} \right. \\
&\qquad \left. \cdot\, e^{-2b\xi}\phi_k^-(x)\phi_l^+(\xi)\phi_m^-(x)\phi_n^+(\xi) \right] d\xi\, dx \qquad (3.7) \\
&= \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} B_{kl}B_{mn}\, e^{-(\lambda_k+\lambda_l+\lambda_m+\lambda_n)T} \\
&\qquad \cdot \left[\int_0^1 e^{-2b\xi}\phi_l^+(\xi)\phi_n^+(\xi)\, d\xi \right] \left[\int_0^1 \phi_k^-(x)\phi_m^-(x)\, dx \right].
\end{aligned}$$

From (2.17), together with the bound

$$|\phi_n^{\pm}(x)| \leqq 2^{1/2}\, e^{\pm bx/2}, \qquad 0 \leqq x \leqq 1, \quad n \geqq 1, \tag{3.8}$$

it follows that

$$|B_{mn}| \leqq 2b^{-1}(e^b-1), \qquad m, n \geqq 1. \tag{3.9}$$

Use of (3.8) and (3.9) in (3.7) gives

$$\begin{aligned}
\|\mathscr{L}_K\|_2^2 &\leqq (2/b)^4(e^b-1)^4\, e^{-2b} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} e^{-(\lambda_k+\lambda_l+\lambda_m+\lambda_n)T} \\
&= (2/b)^4(e^b-1)^4\, e^{-b(2+T)} \left[\sum_{k=1}^{\infty} e^{-(\pi^2 T/b)k^2} \right]^4.
\end{aligned}$$

Upon majorizing the series on the right-hand side by the appropriate definite integral, we obtain

$$\|\mathscr{L}_K\|_2^2 < \Gamma(b, T). \tag{3.10}$$

If $\Gamma(b, T) \leqq 1$, then $\|\mathscr{L}_K\|_2 < 1$, thereby assuring the existence and uniqueness of a time-periodic solution. Q.E.D.

Remark. If any norm of $\mathscr{L}_K$ is less than unity, then an iterative scheme starting from any initial distribution $w^{(0)}(x)$, namely

$$w^{(k+1)}(x) = \int_0^1 K(x;\xi)w^{(k)}(\xi)\, d\xi + Q(x), \tag{3.11}$$

converges in the corresponding function norm to the solution of (2.43). This iteration corresponds to the actual temporal evolution of the probability density as recorded by a stroboscopic device which records images only at the discrete times $t=0$, $2T, 4T, \cdots$. In particular, the convergence of the iterative scheme assures that the solution to problem $\mathcal{P}$, starting from any arbitrary initial distribution, approaches the time-periodic solution in the pertinent function norm as time $t\to\infty$.

Figure 1 summarizes the results of this section. It depicts the two curves respectively defined by

$$T=2b^2/(b^2+4\pi^2) \quad \text{(curve I)},$$

$$\Gamma(b, T)=1 \quad \text{(curve II)}.$$

We have established that there exists a unique time-periodic solution of problem $\mathcal{P}$ for (b, T) lying above curve I, on or above curve II, or both (corresponding to the shaded region in Fig. 1).

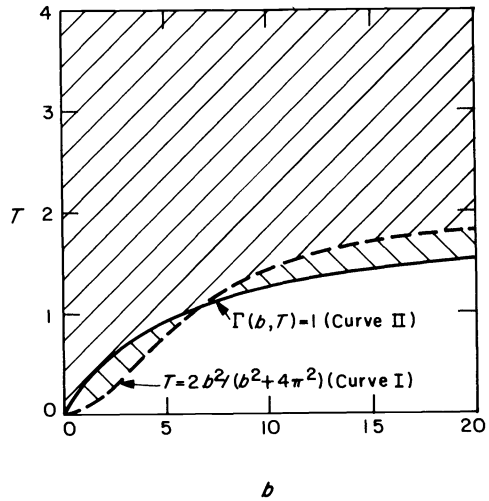


FIG. 1. *(b, T)-plane. Existence and uniqueness of a time-periodic solution of problem $\mathcal{P}$ are assured for (b, T) lying in the shaded region.*

Remark. By deriving a bound on the L^2 norm of the infinite matrix operator C, it can be shown that problem $\mathcal{P}$ possesses a unique time-periodic solution if b and T satisfy the condition

$$\Omega(b, T) \overset{\text{def}}{=} (e^b-1)^2(e^{2b}-1)(2\pi bT)^{-3/2}\, e^{-b(2+T)} \leqq 1.$$

The level curve given by $\Omega(b, T)=1$ is quantitatively similar to curve II. It is not shown in Fig. 1 because its inclusion would not enlarge the shaded region.

4. Regularity of time-periodic solutions. Conditions under which square integrable time-periodic solutions of problem $\mathcal{P}$ exist were established in § 3. We now briefly sketch the argument that leads to a much stronger regularity property of such solutions. The key observation is that at each time t, a time-periodic solution $P(x, t)$ possesses a series representation in which the coefficients multiplying the eigenfunctions $\phi_n^+(x)$ or $\phi_n^-(x)$ $(n\geqq 1)$ can be bounded by $[\mathcal{C}(b, T)\exp(-b\tau/4)]\exp[-(\pi^2\tau/b)n]$, where

$\tau > 0$.[4] This statement applies even at the flipping times. For example, of the two series representations $\sum_{n=0}^{\infty} a_n \phi_n^+(x)$ and $\sum_{n=0}^{\infty} b_n \exp(-\lambda_n T)\phi_n^-(x)$ for $P(x, 0) = P(x, 2T)$, the latter is the appropriate one to consider.

It is well known for trigonometric Fourier series that if the Fourier coefficients can be bounded by $A\vartheta^n$ (with $0 < \vartheta < 1$), then the series represents a function that is analytic in x on $[-\pi, \pi]$ (see, for example, Bary [1, pp. 82–83]). Minor modifications of Bary's proof establish the analogous result for the orthonormal sequences $\{\phi_n^+(x)\}$, $\{\phi_n^-(x)\}$ on the interval $[0, 1]$. With $A = \mathscr{C}(b, T)\exp(-b\tau/4)$ and $\vartheta = \exp(-\pi^2\tau/b)$ we obtain the following theorem.

THEOREM 4.1. *Any square integrable time-periodic solution of problem $\mathscr{P}$ is, in fact, an analytic function of x for all $x \in [0, 1]$ at each fixed time t.*

5. Further study of the infinite linear system (2.31). It has already been established that the eigenfunction expansion coefficients a_n $(n \geqq 1)$ of a possible recurring initial distribution must satisfy (2.31). We now exhibit a simpler infinite matrix equation for the expansion coefficients, and use it to establish a symmetry property of unique time-periodic solutions of problem $\mathscr{P}$.

THEOREM 5.1. *Provided that unity is not an eigenvalue of C, the unique solution of* (2.31) *is also determined uniquely by the simpler system*

$$a_m = \sum_{n=1}^{\infty} E_{mn} a_n + S_m a_0, \qquad m \geqq 1, \tag{5.1}$$

where

$$E_{mn} = (-1)^n e^{b/2} A_{mn} e^{-\lambda_n T} \qquad (m, n \geqq 1), \tag{5.2}$$

$$S_m = e^{b/2} A_{m0} \qquad (m \geqq 1). \tag{5.3}$$

Proof. Using (5.2) and (5.3), together with bounds that utilize (3.8) and (2.11a) in the Parseval identity corresponding to (2.14), it can be shown that $E = [E_{mn}]$ and $S = [S_m]$ are square summable. Thus, Fredholm theory applies to (5.1). Next, we note the equalities

$$C_{mn} = \sum_{k=1}^{\infty} E_{mk} E_{kn}, \tag{5.4}$$

$$R_m = S_m + \sum_{k=1}^{\infty} E_{mk} S_k, \tag{5.5}$$

which follow from (5.2), (5.3), (2.32), (2.33) and (2.18). The hypothesis that unity is not an eigenvalue of C, in conjunction with (5.4), leads to the conclusion that unity also fails to be an eigenvalue of E. Thus, (5.1) possesses a unique solution. Moreover, any solution of (5.1) also satisfies (2.31), as can be verified using (5.1), (5.4) and (5.5). Thus, the unique solutions of (5.1) and (2.31) are, in fact, one and the same. Q.E.D.

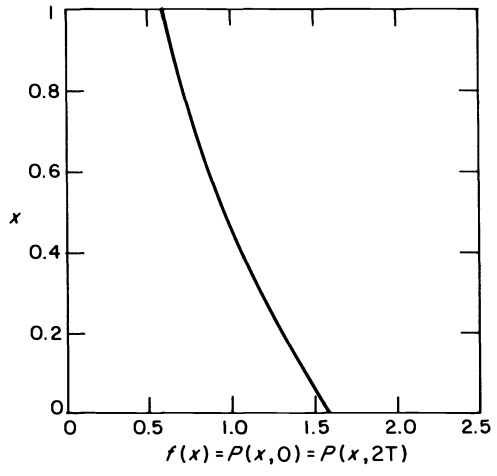
By considering the simpler system (5.1) we can establish the following time-shifted symmetry property of unique time-periodic solutions.

THEOREM 5.2. *Unique time-periodic solutions $P(x, t)$ of problem $\mathscr{P}$ satisfy*

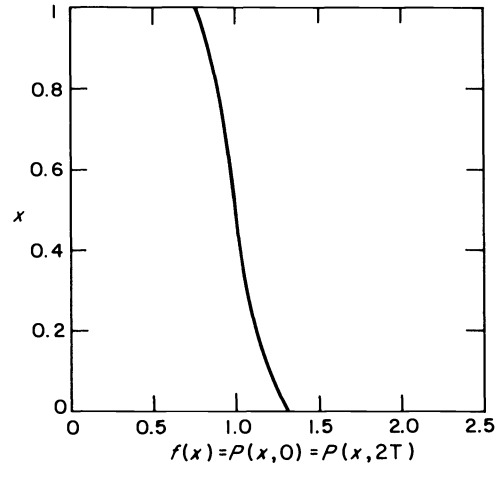
$$P(1 - x, t + T) = P(x, t). \tag{5.6}$$

---

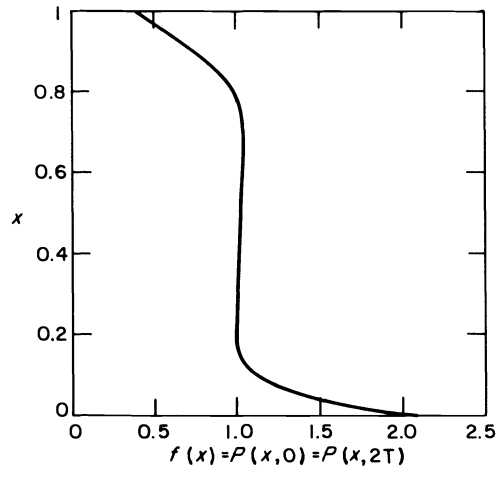
[4] We establish this bound by considering (2.22) for $0 < t \leqq T$ with $\tau = t$, and (2.26) for $T < t \leqq 2T$ with $\tau = t - T$. We then jointly utilize (2.10), the inequality $\exp[-(\pi^2\tau/b)n^2] \leqq \exp[-(\pi^2\tau/b)n]$ for $n \geqq 1$, and the fact that the a_n and b_n are certainly bounded, say by $\mathscr{C}(b, T)$.



(a)



(b)



(c)

FIG. 2. *Numerically generated recurring initial distributions*: (a) $b = 1$, $T = 1$; (b) $b = 1$, $T = 0.1$; (c) $b = 10$, $T = 0.1$.

Proof. We first derive a simple relation between the coefficients a_k and b_k for a unique time-periodic solution. By utilizing (2.29), (2.18), (2.19c), (5.2) and (5.3) we obtain

$$b_0 = e^{b/2} a_0, \quad b_k = (-1)^k e^{b/2} \left[\sum_{n=1}^{\infty} E_{kn} a_n + S_k a_0 \right] \quad \text{for } k \geqq 1. \tag{5.7}$$

For a unique time-periodic solution the coefficients a_n $(n \geqq 1)$ satisfy (5.1). Thus,

$$b_k = (-1)^k e^{b/2} a_k \tag{5.8}$$

for all $k \geqq 0$. This relation, together with (2.13), yields

$$b_k \phi_k^-(x) = a_k \phi_k^+(1-x). \tag{5.9}$$

From (2.22), (2.26) and (5.9) we conclude that a unique time-periodic solution $P(x, t)$ of problem $\mathcal{P}$ can be represented as follows:

$$P(x, t) = \begin{cases} \sum_{m=0}^{\infty} a_m e^{-\lambda_m t} \phi_m^+(x), & 0 \leqq t \leqq T, \\ \sum_{m=0}^{\infty} a_m e^{-\lambda_m (t-T)} \phi_m^+(1-x), & T \leqq t \leqq 2T. \end{cases} \tag{5.10}$$

The theorem is an immediate consequence of (5.10). Q.E.D.

Remark. It can be deduced from the preceding theorem that, in a laboratory reference frame, the probability density distributions for comparable times after each flip are all identical. This is in accordance with physical intuition.

6. Concluding remarks. In order to provide perspective for the appearance of time-periodic solutions of problem $\mathcal{P}$, we present in Figs. 2(a)-(c) examples of recurring initial distributions, generated numerically by solving finite sections of (5.1) for specific parameter values. These spatial distributions also represent the probability density at time $t = 2T$, confirming the expectation of an accumulation of probability density ("sediment") at the bottom of the container after the system has been left undisturbed for one half period. Figures 2(a) and 2(b) show the accumulation to be more pronounced for a larger than for a smaller value of T, again conforming to expectations, since the Brownian particle(s) will then have had more time in which to settle between successive overturnings.

Existence and uniqueness of time-periodic solutions have been proven for a certain region of the (b, T)-plane. One direction for future study is that of determining the eigenvalues and corresponding eigenvectors of the infinite matrix C (or E)—in particular, the first eigenvalue. This would furnish further information regarding conditions under which existence and uniqueness are assured.

REFERENCES

[1] N. K. BARY, *A Treatise on Trigonometric Series*, Vol. I (Authorized translation by M. F. Mullins), Pergamon Press, Oxford, 1964.

[2] J. R. CANNON, *The One-Dimensional Heat Equation*, Addison-Wesley, Reading, MA, 1984.

[3] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Vol.* I, Wiley-Interscience, New York, 1953.

[4] A. M. J. DAVIS AND H. BRENNER, *Steady rotation of a tethered sphere at small, non-zero Reynolds and Taylor numbers: wake interference effects on drag*, J. Fluid Mech., 168 (1986), pp. 151-167.

[5] L. H. DILL AND H. BRENNER, *Taylor dispersion in systems of sedimenting nonspherical Brownian particles* III. *Time-periodic forces*, J. Colloid Interface Sci., 94 (1983), pp. 430-450.

[6] S. J. FARLOW, *An existence theorem for periodic solutions of a parabolic boundary value problem of the second kind*, SIAM J. Appl. Math., 16 (1968), pp. 1223-1226.

[7] ———, *Periodic solutions of nonlinear boundary value problems of the second kind*, Portugal. Math., 32 (1973), pp. 25-37.

[8] E. HELLINGER AND O. TOEPLITZ, *Integralgleichungen und Gleichungen mit unendlichvielen Unbekannten*, in Encyklopädie der Mathematischen Wissenschaften, 2.3.2, H. Burkhardt, W. Wirtinger, R. Fricke and E. Hilb, eds., B. G. Teubner, Leipzig, 1923-27, pp. 1335-1597.

[9] J. HEUSS, *Existenzsätze für periodische Lösungen parabolischer Randwertprobleme mit der Linienmethode*, Ph.D. Dissertation, University of Karlsruhe, Karlsruhe, West Germany, 1979.

[10] D. HILBERT, *Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen*, Chelsea, New York, 1953.

[11] B. KAWOHL AND R. RÜHL, *Periodic solutions of nonlinear heat equations under discontinuous boundary conditions*, Equadiff 82 (Würzburg, 1982), pp. 322-327, Lecture Notes in Mathematics 1017, Springer-Verlag, New York, 1983.

[12] H. KNOLLE, *Periodische Lösungen der 3. Randwertaufgabe einer quasilinearen parabolischen Differentialgleichung*, Ph.D. Dissertation, Gesamthochschule Paderborn, 1977.

[13] A. P. MAL'TSEV, *Convergence and stability of Rothe's method for periodic solution of a quasilinear parabolic equation with nonlinear boundary conditions*, Izv. Vyss. Učebn. Zaved. Radiofizika, 12 (1969), pp. 415-424; Radiophys. and Quantum Electronics, 12 (1969), pp. 331-338.

[13a] ———, *Construction of periodic solutions of boundary value problems for parabolic equations by the method of straight lines*, Izv. Vyss. Učebn. Zaved. Radiofizika, 12 (1969), pp. 1657-1665; Radiophys. and Quantum Electronics, 12 (1969), pp. 1292-1298.

[14] S. G. MIKHLIN, *Linear Integral Equations* (translated from the Russian), Hindustan Publishing, Delhi, India, 1960.

[15] A. NADIM, R. G. COX AND H. BRENNER, *Transport of sedimenting Brownian particles in a rotating Poiseuille flow*, Phys. Fluids, 28 (1985), pp. 3457-3466.

[16] G. H. OTTO AND A. LORENTZ, *Simulation of low gravity conditions by rotation*, AIAA Paper 78-273, AIAA 16th Aerospace Sciences Meeting, Huntsville, Alabama, January 16-18, 1978, 9 pages.

[17] G. PRODI, *Problemi al contorno non lineari per equazioni di tipo parabolico non lineari in due variabili—soluzioni periodiche*, Rend. Sem. Mat. Univ. Padova, 23 (1954), pp. 25-85.

[18] F. RIESZ AND B. SZ.-NAGY, *Functional Analysis* (translated from the 2nd French edition), Frederick Unger, New York, 1971.

[19] I. I. ŠMULEV, *Periodic solutions of boundary value problems without initial conditions for quasilinear parabolic equations*, Dokl. Akad. Nauk. SSSR, 139 (1961), pp. 1318-1321; Soviet Math. Dokl., 2 (1961), pp. 1103-1107.

[20] ———, *Almost-periodic and periodic solutions of a problem for parabolic equations involving an oblique derivative*, Differentsial'nye Uravnenija, 5 (1969), pp. 2225-2236; Differential Equations, 5 (1969), pp. 1668-1676.

[21] O. VEJVODA, *Partial Differential Equations: Time-Periodic Solutions*, Martinus Nijhoff, Boston, 1982.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

013

REMARKS ON THE SECOND EIGENVALUE OF A SYMMETRIC SIMPLY CONNECTED PLANE REGION*

CHAO-LIANG SHEN†

Abstract. In this paper we study the second eigenfunctions of the fixed membrane eigenvalue problem on a symmetric domain. Using the notion of axially symmetric functions we prove a comparison theorem for two eigenvalues when the nodal sets of the corresponding eigenfunctions are known. This result implies that the second eigenvalues of certain membranes are nondegenerate.

Key words. eigenvalues, multiplicity, eigenfunctions, nodal set

AMS(MOS) subject classifications. 35B05, 35P99

1. Introduction. The purpose of this paper is to study the eigenspace of the second eigenvalue of a fixed membrane problem (§ 2), and to find a sufficient condition for the nondegeneracy of the second eigenvalue of a simply connected plane region, symmetric with respect to the x-, y-axes (§ 3). In § 3, using the notion of axially symmetric functions proposed by Payne, Weinberger and Weinstein [4], [5], [6], we prove a comparison theorem (Theorem 7) for two eigenvalues of a membrane when the nodal sets of the corresponding eigenfunctions are known. Using this result, we find that if f is continuous and strictly decreasing in $[0, a]$, $f(0)>0$, $f(a)=0$, but $x^2+[f(x)]^2$ is strictly increasing in $[0, a]$, then the second eigenvalue of the fixed membrane problem on the region $\bar{\Omega} := \{(x, y)\in\mathbb{R}^2: -a\leqq x\leqq a,\ -f(|x|)\leqq y\leqq f(|x|)\}$ is simple (Theorem 8).

2. A decomposition of the eigenspace of the second eigenvalue of a two-symmetric simply connected plane region. Ω shall denote a 2-symmetric simply connected plane region (i.e., Ω is separately symmetric with respect to the x-, y-axes). We shall use the following notation and terminology:

Let λ_2 be the second eigenvalue of the fixed membrane problem on Ω.

(1) If $u\in C^2(\Omega)\cap C(\bar{\Omega})$, $u=0$ on $\partial\Omega$, and $u\not\equiv 0$ such that $\Delta u+\lambda_2 u=0$, we shall call it a *second eigenfunction* of Ω.

(2) $E(\lambda_2)$ denotes the vector space consisting of 0 and second eigenfunctions.

(3) For $u\in E(\lambda_2)$, $u\not\equiv 0$, the set $N(u)=\{(x, y)\in\Omega: u(x, y)=0\}$ is called the *nodal line* (*nodal set*) of u. Any connected component of $\Omega\backslash N(u)$ is called a *nodal domain* of u. It is known that for $u\in E(\lambda_2)\backslash\{0\}$, u has exactly two nodal domains; the signs of u on these two nodal domains are different (see [1], [2]).

(4) $S(\lambda_2)=\{u\in E(\lambda_2)$: $u=0$, or $(0, 0)\in N(u)$ and $N(u)$ is symmetric with respect to $(0, 0)\}$.

(5) $\mathscr{F}(\lambda_2)=\{u\in E(\lambda_2)$: $u(x, y)=u(x, -y)=u(-x, y)$ for all (x, y) in $\Omega\}$.

(6) A function u in Ω is *2-symmetric* if $u(x, y)=u(x, -y)=u(-x, y)$ for all (x, y) in Ω. A subset N of Ω is *2-symmetric* if (x, y) in N implies that $(x, -y)$ and $(-x, y)$ are all in N.

It follows from the maximum principle for subharmonic functions that $N(u)$ cannot have isolated points. It is also known that critical nodal points are isolated (see [1, § 2]). Using these facts we see that for a second eigenfunction u, $N(u)$ is either a simple curve with end points on $\partial\Omega$ or a simple closed loop in Ω.

* Received by the editors March 21, 1986; accepted for publication (in revised form) March 24, 1987. This research was supported in part by National Science Council grant 75-0208-M007-23 of the Republic of China.

† Institute of Mathematics, National Tsing Hua University, Hsinchu, Taiwan 300, Republic of China.

We note that in general if we have a segment L in the nodal set $N(v)$ of an eigenfunction v of Ω, then L can be extended either to a (possibly self-intersected) curve in $N(v)$ with end points on $\partial\Omega$, or to a (possibly self-intersected) loop in $N(v)$. Since when nodal lines meet at a point they form an equiangular system with an even number of angles ([1, Thm. 2.5], [3, §§ 25.12, 25.13]), if we have an equiangular system in $N(v)$ at a nodal point P with $2j_0$ angles, then the eigenfunction v has at least j_0+1 nodal domains.

LEMMA 1. *For $u_1, u_2 \in E(\lambda_2)$, if $N(u_1) = N(u_2)$, then $u_1 = \alpha u_2$ for some $\alpha \in \mathbb{R}$.*

Proof. We know that any second eigenfunction has exactly two nodal domains. Let Ω_1 and Ω_2 be the nodal domains of u_1 and u_2. Then $u_i|_{\Omega_j}$ is a first eigenfunction of Ω_j (see [2]). Thus there exist real numbers α and β such that

$$u_1 = \alpha u_2 \quad \text{in } \Omega_1, \qquad u_1 = \beta u_2 \quad \text{in } \Omega_2.$$

Suppose $P \in \partial\Omega_1 \cap \partial\Omega_2 \cap \Omega$. Then

$$\frac{\partial u_1}{\partial \xi}(P) = \alpha \frac{\partial u_2}{\partial \xi}(P), \qquad \frac{\partial u_1}{\partial \xi}(P) = \beta \frac{\partial u_2}{\partial \xi}(P),$$

for any P, in any direction ξ. Since critical nodal points are isolated, there exist a nodal point P and a direction ξ such that $(\partial u_2/\partial \xi)(P) \neq 0$. Hence $\alpha = \beta$. □

THEOREM 2. *$S(\lambda_2)$ is a vector space.* $\dim_{\mathbb{R}} S(\lambda_2) \leqq 2$.

Proof. For $u \neq 0$ in $S(\lambda_2)$ let Ω_1 and Ω_2 be its nodal domains. Then Ω_1 and Ω_2 are congruent from the setting of $S(\lambda_2)$, and one is the reflection with respect to $(0, 0)$ of the other. Therefore it follows immediately from Lemma 1 that $u(x, y) = -u(-x, -y)$ for all (x, y) in Ω. This implies $S(\lambda_2)$ is a vector space.

Choose u_1, u_2 from $S(\lambda_2)\backslash\{0\}$ such that $u_1(x, y) \not\equiv u_1(x, -y)$ and $u_2(x, y) \not\equiv -u_2(x, -y)$. If no such u_1 exists, then for all u in $S(\lambda_2)$, $u(x, y) = u(x, -y) = -u(-x, -y)$ implies that $N(u)$ contains the y-axis portion (and hence is the y-axis portion) of Ω. Then Lemma 1 implies that $\dim S(\lambda_2) \leqq 1$. Similarly if no such u_2 exists, then $\dim S(\lambda_2) \leqq 1$. Assume both u_1 and u_2 do exist. Define

$$v_1(x, y) = u_1(x, y) - u_1(x, -y),$$
$$v_2(x, y) = u_2(x, y) + u_2(x, -y).$$

Then $N(v_1)$ (resp., $N(v_2)$) consists of the x-axis portion (resp., the y-axis portion) of Ω. For any u in $S(\lambda_2)$, by applying Lemma 1, we can find real numbers α, β such that $u(x, y) - u(x, -y) = \alpha v_1(x, y)$, $u(x, y) + u(x, -y) = \beta v_2(x, y)$. Thus u belongs to the span of v_1 and v_2. Therefore $\dim S(\lambda_2) \leqq 2$. (We also note that $\int_\Omega v_1 v_2 = 0$.) □

THEOREM 3. $\dim_{\mathbb{R}} \mathcal{F}(\lambda_2) \leqq 1$.

Proof. From its definition it is clear that $\mathcal{F}(\lambda_2)$ is a real vector space. Assume $\dim \mathcal{F}(\lambda_2) \geqq 2$. Then there would exist linearly independent u and v in $\mathcal{F}(\lambda_2)$. Since u and v are 2-symmetric, if $N(u)$ were not a loop in Ω, then, by the symmetry of u, $N(u)$ would be the x-axis or the y-axis portion of Ω. This is a contradiction to the 2-symmetry of u, since u has distinct signs on distinct nodal domains. Thus both $N(u)$ and $N(v)$ are 2-symmetric loops in Ω. Let Ω_u and Ω_v be the simply connected nodal domains of u and v, respectively. We may assume u and v have the same sign in $\Omega_u \cap \Omega_v$ (note that $(0, 0) \in \Omega_u \cap \Omega_v$), then we can find $\alpha \in \mathbb{R}_+$ such that $u(0, 0) - \alpha v(0, 0) = 0$. Since $u - \alpha v \neq 0$, $u - \alpha v$ is in $\mathcal{F}(\lambda_2)$. The previous argument implies that $N(u - \alpha v)$ is a 2-symmetric loop in Ω. Thus $(u - \alpha v)(0, 0) = 0$ leads to a contradiction unless $u - \alpha v = 0$. Hence $\dim_{\mathbb{R}} \mathcal{F}(\lambda_2) \leqq 1$. □

COROLLARY 4. *If $u, v \in E(\lambda_2)$ such that $N(u)$ and $N(v)$ are 2-symmetric simply closed loops in Ω, then there exists a real number α such that $u = \alpha v$, i.e., $N(u) = N(v)$.*

Proof. Since $N(u)$ and $N(v)$ are 2-symmetric loops in Ω, the nodal domains of u and v are 2-symmetric regions. By the 2-symmetry of the nodal domains and the simplicity of the first eigenvalue (recall that the restriction of u to any one of its two nodal domains is an eigenfunction of the first eigenvalue λ_2 of that domain), we see immediately that both u and v are 2-symmetric. Thus $u, v \in \mathcal{F}(\lambda_2)$. Since $\dim \mathcal{F}(\lambda_2) \leqq 1$, we are done. □

It is clear from the definitions of $\mathcal{F}(\lambda_2)$ and $S(\lambda_2)$ that for $u \in \mathcal{F}(\lambda_2)$, $v \in S(\lambda_2)$ we have $\int_\Omega uv = 0$.

THEOREM 5. $E(\lambda_2) = \mathcal{F}(\lambda_2) \oplus S(\lambda_2)$. *In particular* $\dim_{\mathbb{R}} E(\lambda_2) \leqq 3$.

Proof. For $u \in E(\lambda_2)$ define

$$u_A(x, y) = u(x, y) + u(-x, -y),$$

$$u_B(x, y) = u(x, y) - u(-x, -y),$$

and denote $E_A(\lambda_2) = \{u_A : u \in E(\lambda_2)\}$, $E_B(\lambda_2) = \{u_B : u \in E(\lambda_2)\}$. Then E_A is orthogonal to E_B in the L^2 sense. We note that $u_B(0, 0) = 0$, $u_B(-x, -y) = -u_B(x, y)$; thus $u_B \in S(\lambda_2)$. $u_A(x, y) = u_A(-x, -y)$. Next, for $u \in E_A(\lambda_2)$, define

$$u_C(x, y) = u(x, y) - u(x, -y),$$

$$u_D(x, y) = u(x, y) + u(x, -y).$$

Let $E_C(\lambda_2) = \{u_C : u \in E_A(\lambda_2)\}$, $E_D(\lambda_2) = \{u_D : u \in E_A(\lambda_2)\}$. Then $E_C(\lambda_2)$ is contained in $S(\lambda_2)$ (in fact, $E_C(\lambda_2) = \{0\}$). For $w \in E_D(\lambda_2)$, $w(x, y) = w(x, -y) = w(-x, -y)$; i.e., $w \in \mathcal{F}(\lambda_2)$. Summarizing the previous argument, we find that $E(\lambda_2) \subseteq \mathcal{F}(\lambda_2) \oplus S(\lambda_2)$. Thus $E(\lambda_2) = \mathcal{F}(\lambda_2) \oplus S(\lambda_2)$. $\dim_{\mathbb{R}} E(\lambda_2) \leqq 3$ follows immediately from Theorems 2 and 3. □

We note that we can employ the arguments in § 25.13 of [3] and the arguments similar to those given by Cheng in [1, Thm. 3.4] to derive the following multiplicity estimate: *Let $m_\Omega(n)$ denote the multiplicity of the nth eigenvalue λ_n of the fixed membrane problem on a bounded simply connected open region Ω in $\mathbb{R}^2$ with reasonably regular boundary. Then $m_\Omega(n) \leqq n(n+1)/2$.* The idea of the proof, as was pointed out by Cheng, is to show the order of vanishing of an nth eigenfunction u is at most $n-1$. Let P be a nodal point of u. We may assume that $P = (0, 0)$ and expand u in a neighborhood of $(0, 0)$ into the following Fourier-Bessel series:

$$u(r, \theta) = \sum_{m=0}^{\infty} J_m(\sqrt{\lambda_n} r)(a_m \cos m\theta + b_m \sin m\theta) \qquad (b_0 = 0).$$

Define the *order of vanishing* of u at $(0, 0)$ to be the largest integer k such that $u(0, 0) = 0$ and $du(0, 0) = 0, \cdots, d^{k-1}u(0, 0) = 0$. If $u(0, 0) = 0$ and $du(0, 0) = 0, \cdots, d^{n-1}u(0, 0) = 0$, then the coefficients $a_0, b_0, a_1, b_1, \cdots, a_{n-1}, b_{n-1}$ are all zeros. This implies that if j_0 is the first term so that $(a_{j_0}, b_{j_0}) \neq (0, 0)$, then $j_0 \geqq n$. For this j_0, say $a_{j_0} \neq 0$ ($b_{j_0} \neq 0$ is discussed similarly). Then $\cot(j_0\theta) = -b_{j_0}/a_{j_0}$ determines j_0 segments of $N(u)$ which intersect at $(0, 0)$ and form an equiangular system with $2j_0$ angles. This implies there are at least $j_0 + 1$ nodal domains (see the note we give before Lemma 1). $j_0 + 1 > n$, which is a contradiction to Courant's nodal domain theorem. Thus the order of vanishing of an nth eigenfunction of a plane fixed membrane problem is at most $n-1$. Then, by using Cheng's argument (on $\mathbb{R}^2$ the dimension of the space of constant coefficient partial differential operators of order less than or equal to $n-1$ is equal to $n(n+1)/2$), we can obtain $m_\Omega(n) \leqq n(n+1)/2$.

3. Comparing two eigenvalues when the nodal sets of the corresponding eigenfunctions are known. Let Ω be a simply connected plane region symmetric with respect to

the x-axis and the y-axis. Let ϕ, ψ be two eigenfunctions of Ω such that $N(\phi)$ lies on the x-axis and $N(\psi)$ lies on the y-axis. In this section we shall prove a comparison theorem (Theorem 7) which states that under certain conditions on $\partial\Omega$, we can compare the corresponding eigenvalues. Using this theorem we find that the second eigenvalues of certain plane regions are simple (Theorem 8).

THEOREM 6. *Let $\Omega = \Omega_f$ be a simply connected plane region with smooth boundary such that $\bar{\Omega}_f = \{(x, y): -a \leqq x \leqq a,\ -f(|x|) \leqq y \leqq f(|x|)\}$, where f is a continuous strictly decreasing function in $[0, a]$, $f(x) > 0$ in $[0, a)$, $f(a) = 0$, $f(0) = b$. Let g be the inverse function of f. Construct two four-dimensional domains Ω_1, Ω_2 as follows:*

$$\bar{\Omega}_1 = \{(x, y, z, w) \in \mathbb{R}^4: y^2 + z^2 + w^2 \leqq f(|x|)^2,\ -a \leqq x \leqq a\},$$

$$\bar{\Omega}_2 = \{(x, y, z, w) \in \mathbb{R}^4: x^2 + z^2 + w^2 \leqq g(|y|)^2,\ -b \leqq y \leqq b\}.$$

Then $\bar{\Omega}_1 \subseteq \bar{\Omega}_2$ if and only if $x^2 + [f(x)]^2$ is an increasing function in $[0, a]$. Furthermore, if $x^2 + [f(x)]^2$ is strictly increasing, then $\bar{\Omega}_1 \subsetneqq \bar{\Omega}_2$.

Proof. Since both $\bar{\Omega}_1$ and $\bar{\Omega}_2$ are simply connected, $\bar{\Omega}_1 \subseteq \bar{\Omega}_2$ if and only if $\partial\Omega_1 \subseteq \bar{\Omega}_2$. To show that $\bar{\Omega}_1 \subseteq \bar{\Omega}_2$, it is sufficient to prove that if $y^2 + r^2 = f(x)^2$, $(x \geqq 0,\ y \geqq 0)$, then $x^2 + r^2 \leqq g(y)^2$. For x, y, r such that $y^2 + r^2 = f(x)^2$, the inequality $x^2 + r^2 \leqq g(y)^2$ holds if and only if $x^2 + f(x)^2 - y^2 \leqq g(y)^2$; i.e., $x^2 + f(x)^2 \leqq y^2 + g(y)^2$. Since $0 \leqq y \leqq f(x)$ and f is strictly decreasing, $y = f(x_1)$ for some $x \leqq x_1 \leqq a$. Thus $\bar{\Omega}_1 \subseteq \bar{\Omega}_2$ if and only if $x^2 + [f(x)]^2$ is an increasing function. It is clear from the previous argument that if $x^2 + [f(x)]^2$ is strictly increasing, then $\bar{\Omega}_1 \neq \bar{\Omega}_2$. □

Example. Let $a > b > 0$ and let Ω be the ellipse $(x^2/a^2) + (y^2/b^2) < 1$. Then $\Omega = \Omega_f$, where $f = b\sqrt{1 - (x^2/a^2)}$ satisfies the conditions stated in Theorem 6.

Let Ω_f be as in Theorem 6, and let $x^2 + [f(x)]^2$ be strictly increasing in $[0, a]$. Suppose ϕ is an eigenfunction of Ω_f such that $N(\phi)$ lies on the x-axis, ψ is an eigenfunction of Ω_f such that $N(\psi)$ lies on the y-axis, and the corresponding eigenvalues of ϕ, ψ are λ, μ, respectively. Let $\phi = y\nu$. Then $\Delta\phi + \lambda\phi = 0$ implies that $(\partial^2\nu/\partial x^2) + (1/y^2)(\partial/\partial y)(y^2(\partial\nu/\partial y)) + \lambda\nu = 0$. If we denote $V(x, y_1, y_2, y_3) = \nu(x, y)$, where $y^2 = y_1^2 + y_2^2 + y_3^2$, then V is an axially symmetric function on the four-dimensional domain Ω_1 (the x-axis is the axis of symmetry of Ω_1). $(\partial^2/\partial x^2) + (1/y^2)(\partial/\partial y)(y^2(\partial/\partial y))$ is the axially symmetric Laplacian [4], [5], [6]. Since $V > 0$ in Ω_1, λ is the first eigenvalue of the four-dimensional membrane Ω_1 (with fixed boundary). Similarly μ is the first eigenvalue of the four-dimensional membrane Ω_2. Since $\bar{\Omega}_2 \supsetneqq \bar{\Omega}_1$, $\mu < \lambda$. Thus we obtain the following comparison theorem.

THEOREM 7. *Suppose $\Omega = \Omega_f$, f satisfies the conditions stated in Theorem 6, and $x^2 + [f(x)]^2$ is strictly increasing. Suppose ϕ and ψ are eigenfunctions of eigenvalues λ, μ, respectively, such that the nodal set of ϕ lies on the x-axis and the nodal set of ψ lies on the y-axis. Then $\lambda > \mu$.*

THEOREM 8. *Let f, Ω_f be as in Theorem 7. Then the second eigenvalue of the fixed membrane problem on Ω_f is simple; the eigenfunction is an odd function of x, even in y.*

Proof. Since Ω_f is convex in the x-direction and symmetric about the y axis, it follows from Payne's Theorem [4, Thm. I] that the second eigenfunction of Ω_f cannot have an interior closed nodal curve. Thus, by Theorem 5, $\dim E(\lambda_2) \leqq 2$. If $\dim E(\lambda_2) = 2$, then there is a second eigenfunction ϕ whose nodal curve lies on the x-axis, and there is another second eigenfunction ψ whose nodal curve lies on the y-axis; then Theorem 7 implies $\lambda_2 > \lambda_2$, which is absurd. □

Example. Let $\Omega = \{(x, y): (x^2/a^2) + (y^2/b^2) < 1\}$, $a > b$. Then $\Omega = \Omega_f$, where $f = b\sqrt{1 - (x^2/a^2)}$ and f satisfies the condition stated in Theorem 7. Thus the second eigenvalue λ_2 of Ω is simple and the nodal curve of corresponding eigenfunction lies

on the y-axis. Furthermore, Payne showed that $\lambda_2 \geqq [\pi/(4\int_\Omega x^2\,dx\,dy)]^{1/2} j_1^2$, where j_1 is the first positive zero of the Bessel function $J_1(x)$.

REFERENCES

[1] S. Y. CHENG, *Eigenfunctions and nodal sets*, Comment. Math. Helv., 51 (1976), pp. 43–55.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. I, Wiley-Interscience, New York–London, 1953.

[3] A. SOMMERFELD, *Partial Differential Equations in Physics*, Academic Press, New York–London, 1949.

[4] L. E. PAYNE, *On two conjectures in the fixed membrane eigenvalue problem*, Z. Angew. Math. Phys., 24 (1973), pp. 721–728.

[5] L. E. PAYNE AND H. F. WEINBERGER, *A Faber–Krahn inequality for wedge-like membranes*, J. Math. Phys., 39 (1960), pp. 182–188.

[6] A. WEINSTEIN, *Generalized axially symmetrical potential theory*, Bull. Amer. Math. Soc., 59 (1952), pp. 20–38.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

014

REGULARITY OF THE SOLUTION OF ELLIPTIC PROBLEMS WITH PIECEWISE ANALYTIC DATA. PART I. BOUNDARY VALUE PROBLEMS FOR LINEAR ELLIPTIC EQUATION OF SECOND ORDER*

I. BABUŠKA† AND B. Q. GUO‡

Abstract. This paper is the first in a series devoted to the analysis of the regularity of the solution of elliptic partial differential equations with piecewise analytic data. The present paper analyzes the case of linear, second order partial differential equation of elliptic type. It concentrates on the case when the domain $\Omega \subset \mathbf{R}^2$ is a polygon, boundary conditions are of changing type and coefficients are analytic on $\bar{\Omega}$. The main result states that the solution belongs to a countably normed space based on weighted Sobolev spaces of all orders with weights located in the vertices of the domain and at the points where the type of boundary conditions changes.

These results are essential for the design and the analysis of the *h-p* version of the finite element method for solving the elliptic differential equations of structural engineering (see [6], [11], [12]).

Key words. elliptic equation with piecewise analytic data, Dirichlet problem, corner singularities

AMS(MOS) subject classifications. 35B65, 35D10, 35G15, 35J05

1. The preliminaries.

1.1. Introduction. In applications, as for example in structural mechanics, the problems of elliptic partial differential equations are typically characterized by piecewise analytic input data. The boundary of the domain is piecewise analytic with corners and edges; the coefficients of the equation are piecewise analytic with interfaces having corners and edges. The type of boundary condition is abruptly changing but they are piecewise analytic, etc.

The regularity theory is typically developed in the framework of Sobolev spaces. We refer here, for example, to the survey [15] and to the monographs [10], [14] addressing the problem of the unsmooth boundary. We refer to [2] for more classical results. Results mentioned above do not characterize sufficiently accurately the class of solutions of problems of applications. The detailed knowledge of the properties of the solutions of problems of applications is essential for the design and analysis of effective numerical methods for solving these problems. We mention here, for example, the *h-p* version of the finite element method which was recently developed and is very successfully used in practice.[1] For more about the theory and practice of the *h-p* version we refer to [6], [17], [18].

The solution of the problem with piecewise analytic data is analytic with the exception of special areas of the domain, where the solution has singular character. Typically it happens in the neighborhood of the corners of the domain, places where the type of the boundary condition changes, etc.

This paper, which is the first one in a series of papers, deals with the problem of characterizing the regularity of the solution of the linear partial differential equation of elliptic type on a polygonal domain. It addresses the case of constant and analytic

* Received by the editors May 12, 1986; accepted for publication April 28, 1987.

† Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. This research was partially supported by the Office of Naval Research under contract N00014-85-K-0169.

‡ Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. This research was partially supported by the Air Force Office of Scientific Research under grant AFOSR-80-0277.

[1] Program PROBE of Noetic Technologies, St. Louis.

coefficients. The main tool of the characterization of the solution is the theory of countably normed spaces based on weighted Sobolev spaces of all orders, where the weights are placed in the vertices of the domain. The main result is that the solution is from the set $B^2_\beta(\Omega)$ of functions which belong to the weighted Sobolev spaces $H^{k,2}_\beta(\Omega)$ for $k=2,\cdots$, and $\|u\|_{H^{k,2}_\beta(\Omega)} \leqq Cd^{k-2}(k-2)!$ with C and d independent of k. The main theorem of the paper is Theorem 2.1 addressing the case of the Poisson equation and its generalization for the general equation with analytic coefficients is given in Theorem 3.1. Theorem 3.1 can be further generalized for the case when the coefficients have singular behavior in the neighborhood of the corners too. (Problems of this type are important in applications when nonlinear equations are considered.) Section 1 gives basic notation and preliminaries. Section 2 deals with the regularity of the solution of the Poisson problem. Section 3 deals with the general equation and § 4, the Appendix, proves some technical lemmas used in the paper.

1.2. Notation. Throughout this paper we shall denote integers by i, j, k, l, m, n. By $\mathbf{R}^1$ and $\mathbf{R}^2$ we shall denote the one- and two-dimensional Euclidean space. If $Q \subset \mathbf{R}^1$, respectively, $Q \subset \mathbf{R}^2$, then $\bar{Q}$ denotes the closure of Q in $\mathbf{R}^1$, respectively, in $\mathbf{R}^2$.

By Ω we denote the polygonal domain in $\mathbf{R}^2$ with boundary $\partial\Omega = \Gamma$, the vertices A_i, $i=1,\cdots,M$, and Γ_i, $i=1,\cdots,M$ the open edges of $\partial\Omega$ connecting A_i and A_{i+1} ($A_1 = A_{M+1}$). Obviously we have $\partial\Omega = \bigcup_{i=1}^M \bar{\Gamma}_i$. By ω_i we denote the measure of the interior angle of Ω at A_i. We allow also $\omega_i = 2\pi$, and $\omega_i = \pi$ and the polygon Ω has hence to be understood in this generalized sense. Let further $\Gamma = \bar{\Gamma}^0 + \bar{\Gamma}^1$, $\Gamma^0 = \bigcup_{i\in\mathscr{D}} \bar{\Gamma}_i$, $\Gamma^1 = \Gamma - \Gamma^0$ where $\mathscr{D}$ is some subset of set $\{1, 2, \cdots, M\}$. Γ^0 will be sometimes referred to as Dirichlet boundary and Γ^1 as Neumann boundary.

By $H^m(\Omega)$ (resp. $H^m(Q)$), $m \geqq 0$, m integer, we denote the Sobolev space of functions with square integrable derivatives of order $\leqq m$ on Ω (resp. Q) furnished with the norm:

$$\|u\|^2_{H^m(\Omega)} = \sum_{0\leqq|\alpha|\leqq m} \|D^\alpha u\|^2_{L_2(\Omega)},$$

$$\alpha = (\alpha_1, \alpha_2), \quad \alpha_i \geqq 0, \quad \text{integers}, \quad i = 1, 2,$$

$$|\alpha| = \alpha_1 + \alpha_2,$$

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} = u_{x_1^{\alpha_1} x_2^{\alpha_2}}.$$

As usual $H^0(\Omega) = L_2(\Omega)$. Further let

$$H^1_0(\Omega) = \{u \in H^1(\Omega) \,|\, u = 0 \text{ on } \Gamma^0\}$$

and

$$|u|^2_{H^m(\Omega)} = \sum_{|\alpha|=m} \|D^\alpha u\|^2_{H^0(\Omega)},$$

$$|D^m u|^2 = \sum_{|\alpha|=m} |D^\alpha u|^2.$$

By $r_i(x) = |x - A_i|$, $i = 1, \cdots, M$, we shall denote the Euclidean distance between x and the vertex A_i of Ω. Let $\beta = (\beta_1, \beta_2, \cdots, \beta_M)$ be an M-tuple of real numbers, $0 < \beta_i < 1$, $i = 1, \cdots, M$. For any integer k let $\beta \pm k = (\beta_1 \pm k, \cdots, \beta_M \pm k)$. Further we denote $\Phi_\beta(x) = \prod_{i=1}^M r_i^{\beta_i}(x)$ and $\Phi_{\beta\pm k} = \prod_{i=1}^M r_i^{\beta_i \pm k}(x)$.

By $H_\beta^{m,l}(\Omega)$, $m \geqq l \geqq 0$, l an integer ($H_\beta^{m,0}(\Omega) = H_\beta^m(\Omega)$) we denote the completion of the set of all infinitely differentiable functions under the norm

$$\|u\|^2_{H_\beta^{m,l}(\Omega)} = \|u\|^2_{H^{l-1}(\Omega)} + \sum_{\substack{|\alpha|=k \\ k=l}}^{m} \| |D^\alpha u| \Phi_{\beta+k-l} \|^2_{L_2(\Omega)}, \qquad l \geqq 1,$$

$$\|u\|^2_{H_\beta^{m,0}(\Omega)} = \|u\|^2_{H_\beta^m(\Omega)} = \sum_{\substack{|\alpha|=k \\ k=0}}^{m} \| |D^\alpha u| \Phi_{\beta+k} \|^2_{L_2(\Omega)}, \qquad l = 0.$$

For $m = l = 0$ we shall write $H_\beta^{0,0}(\Omega) = L_\beta(\Omega)$. The space $H_\beta^{m,2}(\Omega)$ was introduced and widely used in [5].

For $0 < \delta \leqq \infty$, $0 < \omega \leqq 2\pi$ let

$$S = S_\delta^\omega = \{(r, \theta) \mid 0 < r < \delta, 0 < \theta < \omega\},$$

$$\mathscr{D}^\alpha u = \frac{\partial^{|\alpha|} u}{\partial r^{\alpha_1} \partial \theta^{\alpha_2}} = u_{r^{\alpha_1}\theta^{\alpha_2}}$$

and

$$|\mathscr{D}^m u|^2 = \sum_{|\alpha|=m} |r^{-\alpha_2} \mathscr{D}^\alpha u|^2.$$

For $0 < \beta < 1$, $m \geqq l \geqq 1$

$$\mathscr{H}_\beta^{m,l}(S) = \left\{ u \,\middle|\, \|u\|^2_{H^{l-1}(S)} + \sum_{l \leqq |\alpha| \leqq m} \|r^{\alpha_1 - l + \beta} \mathscr{D}^\alpha u\|^2_{L_2(S)} = \|u\|^2_{\mathscr{H}_\beta^{m,l}(S)} < \infty \right\},$$

and $m \geqq 0$

$$\mathscr{H}_\beta^{m,0}(S) = \left\{ u \,\middle|\, \sum_{0 \leqq |\alpha| \leqq m} \|r^{\alpha_1 + \beta} \mathscr{D}^\alpha u\|^2_{L_2(S)} = \|u\|^2_{\mathscr{H}_\beta^{m,0}(S)} < \infty \right\}.$$

Obviously

$$\mathscr{H}_\beta^{0,0}(S) = H_\beta^{0,0}(S) = \mathscr{L}_\beta(S)$$

and

$$\mathscr{H}_\beta^{1,1}(S) = H_\beta^{1,1}(S).$$

We will now show that $\mathscr{H}_\beta^{2,2}(S) = H_\beta^{2,2}(S)$.

LEMMA 1.1. *Let $0 < \delta < \infty$. Then the spaces $\mathscr{H}_\beta^{2,2}(S)$ and $H_\beta^{2,2}(S)$ with $\Phi_\beta = r^\beta$ are equivalent.*

Proof. Observe that

$$u_{x_1} = u_r \cos\theta - u_\theta \frac{\sin\theta}{r},$$

$$u_{x_1 x_1} = u_{rr} \cos^2\theta - u_{r\theta} \frac{\sin 2\theta}{r} + \frac{1}{r^2} u_{\theta^2} \sin^2\theta$$

$$+ \frac{1}{r} u_r \sin^2\theta + \frac{1}{r^2} u_\theta \sin 2\theta.$$

Hence

$$\|u_{x_1 x_1}\|_{\mathscr{L}_\beta(S)} \leqq \sum_{|\alpha|=2} \|r^{\alpha_1 - 2} \mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S)} + \sum_{|\alpha|=1} \|r^{\alpha_1 - 2} \mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S)}.$$

By Lemma A.2 (see Appendix) we have for $|\alpha|=1$

$$\|r^{\alpha_1-2}\mathscr{D}^{\alpha}u\|_{\mathscr{L}_\beta(S)} \leqq C(\delta)\left[\sum_{|\alpha'|=2}\|r^{\alpha_1'-2}\mathscr{D}^{\alpha'}u\|_{\mathscr{L}_\beta(S)}+\|u\|_{H^1(S)}\right]$$

and hence

$$\|u_{x_1x_1}\|_{\mathscr{L}_\beta(S)} \leqq C\|u\|_{\mathscr{H}^{2,2}_\beta(S)}.$$

Similarly, we have the same relation for $u_{x_1x_2}$ and $u_{x_2x_2}$, and hence $H^{2,2}_\beta(S)\subset\mathscr{H}^{2,2}_\beta(S)$. The other direction follows directly. □

Later we will investigate the case S^ω_δ when $\delta=\infty$. In this case we will write Q instead of S^ω_δ.

LEMMA 1.2. *Let Ω be the polygon, then for $j=0, 1$ we have*

$$\int_\Omega \Phi^2_{-1+\beta}|u_{x_1^{1-j}x_2^j}|^2\,dx_1\,dx_2 \leqq C\|u\|^2_{H^{2,2}_\beta(\Omega)}, \tag{1.1a}$$

$$\int_\Omega \Phi^2_{-1+\beta}r_i^{-2j}|u_{r_i^{1-j}\theta_i^j}|^2 r_i\,dr_i\,d\theta_i \leqq C\|u\|^2_{H^{2,2}_\beta(\Omega)}, \tag{1.1b}$$

where (r_i, θ_i) are polar coordinates with respect to A_i, $1\leqq i\leqq M$.

Proof. We can write

$$\Omega=\bigcup_{i=1}^{M} S^{\omega_i}_{\delta_i}(A_i)\cup R,$$

$$R=\Omega-\sum_{i=1}^{M} S^{\omega_i}_{\delta_i/2}(A_i)$$

where $S^{\omega_i}_{\delta_i}(A_i)\subset\Omega$ are sectors with the origin in A_i such that $S^{\omega_i}_{\delta_i}(A_i)\cap S^{\omega_i}_{\delta_j}(A_j)=\varnothing$ for $i\neq j$ and ω_i is the interior angle at A_i. Obviously (1.1a, b) hold on R. Lemma A.3 yields (1.1a) on $S^{\omega_i}_{\delta_i}(A_i)$, Lemma A.2 and Lemma 1.1 yield (1.1b) on $S^{\omega_i}_{\delta_i}(A_i)$. □

We also recall the spaces $W^k_\beta(S)$ introduced by Kondrat'ev (see [14], [15])

$$W^k_\beta(S)=\left\{u\,\middle|\,\sum_{0\leqq|\alpha|\leqq k}\|r^{\beta-k+\alpha_1}\mathscr{D}^{\alpha}u\|_{L_2(S)}=\|u\|^2_{W^k_\beta(S)}<\infty\right\}.$$

Finally let

$$D=\{\tau,\theta\,|\,-\infty<\tau<\infty, 0<\theta<\omega\}$$

and for $h>0$ and $k\geqq 0$ an integer define

$$\mathscr{H}^k_h(D)=\left\{u\,\middle|\,\sum_{0\leqq|\alpha|\leqq k}\int_D e^{2h\tau}|D^\alpha u|^2\,d\tau\,d\theta=\|u\|^2_{\mathscr{H}^k_h(D)}<\infty\right\}.$$

We will write also $\mathscr{H}^0_h(F)=\mathscr{L}_h(D)$.

1.3. The space $\psi^l_\beta(\Omega)$ and $B^l_\beta(\Omega)$. For l an integer $0\leqq l\leqq 2$ let

$$\psi^l_\beta(\Omega)=\{u(x)\,|\,u\in H^{m,l}_\beta(\Omega), m\geqq l\} \tag{1.2}$$

and

$$B^l_\beta(\Omega)=\{u(x)\,|\,u\in\psi^l_\beta(\Omega), \||D^\alpha u|\Phi_{\beta+k-l}\|_{L_2(\Omega)}\leqq Cd^{k-l}(k-l)! \text{ for } |\alpha|=k=l, l+1,\cdots, d\geqq 1, C \text{ independent of } k\}. \tag{1.3}$$

For $l=0$ we shall write $B_\beta(\Omega)$ instead $B^0_\beta(\Omega)$. Constants C and d in (1.3) depend on u.

The space $B^l_\beta(\Omega)$ was defined in Cartesian coordinates. There is also an equivalent definition of $B^l_\beta(\Omega)$ in polar coordinates.

Let (r_i, θ_i) be the polar coordinates with respect to A_i, $i = 1, \cdots, M$ as before.

THEOREM 1.1. *Let* $0 \leqq l \leqq 2$. *Then*

$$\| |D^\alpha u| \Phi_{k-l+\beta} \|_{L_2(\Omega)} \leqq C d^{k-l}(k-l)!, \qquad |\alpha| = k,\ k \geqq l \tag{1.4}$$

if and only if

$$\left(\int_\Omega |\mathscr{D}_i^{\alpha'} u|^2 r_i^{-2\alpha'_2} \Phi^2_{k'-l+\beta} r_i \, dr_i \, d\theta_i \right)^{1/2} \leqq C_i d_i^{k'-l}(k'-l)! \tag{1.5}$$

holds for all $l \leqq |\alpha'| = k' \leqq k$, $\alpha' = (\alpha'_1, \alpha'_2)$ *and* $i = 1, \cdots, M$. *By* $\mathscr{D}_i^\alpha u$ *we denoted differentiation with respect to the polar coordinates* (r_i, θ_i).

Proof. We first prove that if (1.4) holds then (1.5) holds for every $i = 1, \cdots, M$. To this end we fix i and will omit writing the index i. Then

$$u_{r^k} = \sum_{j=0}^{k} \binom{k}{j} u_{x_1^{k-j} x_2^j} \cos^{k-j}\theta \sin^j \theta. \tag{1.6}$$

Hence

$$\begin{aligned} \left(\int_\Omega |u_{r^k}|^2 \Phi^2_{k-l+\beta} r \, dr \, d\theta \right)^{1/2} &\leqq \sum_{j=0}^{k} \binom{k}{j} \| u_{x_1^{k-j} x_2^j} \Phi_{k-l+\beta} \|_{L_2(\Omega)} \\ &\leqq C_1 2^k d^{k-l}(k-l)! \\ &= C_2 (2d)^{k-l}(k-l)! \end{aligned} \tag{1.7}$$

and (1.5) is proven for $\alpha'_2 = 0$.

We show now by induction that for any $k \geqq 1$:

$$u_{\theta^k} = \sum_{m=1}^{k} r^m \sum_{j=0}^{m} \sum_{\substack{l_1, l_2 \geqq 0 \\ l_1 + l_2 = m}} a^{(k)}_{m,j,l_1,l_2} \sin^{l_1}\theta \cos^{l_2}\theta \, u_{x_1^{m-j} x_2^j}, \tag{1.8a}$$

$$A^{(k)}_m = \sum_{j=0}^{m} \sum_{\substack{l_1, l_2 \geqq 0 \\ l_1 + l_2 = m}} |a^{(k)}_{m,j,l_1,l_2}| \leqq 4^k \frac{k!}{m!}. \tag{1.8b}$$

Suppose that (1.8) holds for $k = n-1$. Then

$$\begin{aligned} u_{\theta^n} = \sum_{m=1}^{n-1} r^m \sum_{j=0}^{m} \sum_{\substack{l_1, l_2 \geqq 0 \\ l_1 + l_2 = m}} a^{(n-1)}_{m,j,l_1,l_2} \\ \cdot [-r \sin^{l_1+1}\theta \cos^{l_2}\theta \, u_{x_1^{m-j+1} x_2^j} + r \sin^{l_1}\theta \cos^{l_2+1}\theta \, u_{x_1^{m-j} x_2^{j+1}} \\ + (l_1 \sin^{l_1-1}\theta \cos^{l_2+1}\theta - l_2 \sin^{l_1+1}\theta \cos^{l_2-1}\theta) u_{x_1^{m-j} x_2^j}]. \end{aligned}$$

Comparing the coefficients we get

$$\begin{aligned} a^{(n)}_{m,j,l_1,l_2} = -a^{(n-1)}_{m-1,j,l_1-1,l_2} + a^{(n-1)}_{m-1,j-1,l_1,l_2-1} \\ + (l_1+2) a^{(n-1)}_{m,j,l_1+1,l_2-1} - (l_2+2) a^{(n-1)}_{m,j,l_1-1,l_2+1}. \end{aligned}$$

Thus

$$A^{(n)}_m \leqq 2(1+m) A^{(n-1)}_m + 2A^{(n-1)}_{m-1}.$$

Using the induction assumption we get for $n = k$ and $m \leqq k$

$$A^k_m \leqq 4^k \frac{k!}{m!}$$

and (1.8b) is proven.

Let $D = \max(1, \operatorname{diam}\Omega)$. Then for $k \geqq 1$, $0 \leqq l \leqq 1$

$$\left(\int_\Omega r^{-2k} |u_{\theta^k}|^2 \Phi^2_{k-l+\beta} r\, dr\, d\theta\right)^{1/2}$$

$$\leqq \sum_{m=1}^{k} \sum_{j=0}^{m} \sum_{\substack{l_1,l_2 \geqq 0 \\ l_1+l_2=m}}^{m} |a^{(k)}_{m,j,l_1,l_2}| D^{(k+1-l)(M-1)} \|\Phi_{m-l+\beta} u_{x_1^{m-j} x_2^j}\|_{L_2(\Omega)}$$

$$\leqq C D_1^{k-l} \sum_{m=1}^{k} A^{(k)}_m d^{m-l} (m-l)! \tag{1.9}$$

$$\leqq C D_2^{k-l} \sum_{m=1}^{k} 4^k \frac{k!}{m!} (m-l)!$$

$$\leqq C D_3^{k-l} (k-l)!$$

where C and D_3 are independent of k.

By Lemma 1.2 we have for $j = 0, 1$

$$\int_\Omega \Phi^2_{-1+\beta} |u_{x_1^{1-j} x_2^j}|^2\, dx \leqq C \|u\|^2_{H^{2,2}_\beta(\Omega)} \leqq C. \tag{1.10}$$

Hence (1.9) holds for all $k \geqq l$, $l \leqq 2$, and (1.5) holds for $\alpha_1' = 0$. Combining the arguments we have used above we get (1.5) in full generality.

(2) We will now show that if (1.5) holds for every i, then (1.4) holds too. First we will show that for any $k \geqq 1$

$$u_{x_1^k} = \sum_{m=1}^{k} \sum_{j=0}^{m} \sum_{\substack{l_1,l_2 \geqq 0 \\ l_1+l_2=k}} b^{(k)}_{m,j,l_1,l_2} \sin^{l_1}\theta \cos^{l_2}\theta\, r^{-(k-m+j)} u_{r^{m-j}\theta^j}, \tag{1.11a}$$

$$B^{(k)}_m = \sum_{j=0}^{m} \sum_{\substack{l_1,l_2 \geqq 0 \\ l_1+l_2=k}} |b^{(k)}_{m,j,l_1,l_2}| \leqq 5^k \frac{k!}{m!}. \tag{1.11b}$$

It is easy to check that (1.11) holds for $k = 0, 1$. Analogously as in the first part we get

$$b^{(k)}_{m,j,l_1,l_2} = b^{(k-1)}_{m-1,j,l_1,l_2-1} - b^{(k-1)}_{m-1,j-1,l_1-1,l_2}$$
$$-(l_1+2) b^{(k-1)}_{m,j,l_1,l_2-1} + (l_2+2) b^{(k-1)}_{m,j,l_1-1,l_2} - (k-m+j) b^{(k-1)}_{m,j,l_1,l_2-1}$$

and hence

$$B^{(k)}_m \leqq 2B^{(k-1)}_{m-1} + 2kB^{(k-1)}_m + kB^{(k-1)}_m \leqq 2B^{(k-1)}_{m-1} + 3kB^{(k-1)}_m.$$

Using the induction hypothesis we get (1.11b). Using (1.11) we get for $k \geqq l$ and $l \geqq 1$

$$\left(\int_\Omega |u_{x_1^k}|^2 \Phi^2_{k-l+\beta}\, dx\right)^{1/2} \leqq C D^{(k-l)} \sum_{m=1}^{k} B^{(k)}_m d^{m-l} (m-l)!$$
$$\leqq C D_2^{(k-l)} (k-l)! \tag{1.12}$$

where D_2 and C are independent of k. For $j = 0, 1$ we have by Lemma 1.2

$$\int_\Omega \Phi^2_{-1+\beta} r_i^{-2j} |u_{r_i^{1-j}\theta_i^j}|^2 r_i\, dr_i\, d\theta_i \leqq C \|u\|^2_{H^{2,2}_\beta(\Omega)} \leqq \tilde{C}.$$

Hence (1.12) holds for $0 \leqq l \leqq 2$, $k \geqq l$ and (1.4) holds for $\alpha_2 = 0$. The general case can be proven quite analogously. □

Theorem 1.1 yields an equivalent definition of $B^l_\beta(\Omega)$, $0 \leqq l \leqq 2$

$$B^l_\beta(\Omega) = \left\{ u \in \psi^l_\beta(\Omega) \left| \left(\int_\Omega r_i^{-2\alpha_2} \Phi^2_{k-l+\beta} |\mathscr{D}_i^\alpha u|^2 r_i \, dr_i \, d\theta_i \right)^{1/2} \leqq C d^{k-l} (k-l)! \right. \right. \tag{1.13}$$
$$\left. \text{for any } k \geqq l \text{ and } |\alpha| = k;\ C \text{ and } d \text{ independent of } k,\ i = 1, \cdots, M \right\}$$

where (r_i, θ_i) are polar coordinates with the origin in A_i and $\mathscr{D}_i^\alpha u = u_{r_i^{\alpha_1} \theta_i^{\alpha_2}}$. In what will follow both definitions will be used interchangeably.

Remark 1. Let $S = \{r, \theta \,|\, 0 < r < 1, 0 < \theta < \omega\}$. Then $u_1 = r^\alpha \sin\theta$ and $u_2 = r^\alpha \operatorname{lgr} \sin\theta$, $0 < \alpha < 1$ belong to $B^2_\beta(S)$, $\beta \in (1-\alpha, 1)$ but not to $B^2_{1-\alpha}(S)$.

1.4. The spaces $H^{m-1/2}(\gamma)$ and $H^{m-1/2,l-1/2}_\beta(\gamma)$. Let $Q \subset R^2$ be an open bounded set with a piecewise analytic boundary ∂Q and let γ be part of, or the whole boundary ∂Q. We define $H^{m-1/2}(\gamma)$, $m \geqq 1$ as the set of all functions φ on γ such that there exists $f \in H^m(\gamma)$, with $\varphi = f|_\gamma$. The norm is defined by

$$\|\varphi\|_{H^{m-1/2}(\gamma)} = \inf \|f\|_{H^m(Q)}$$

where the infimum is taken over all functions $f \in H^m(Q)$ with $f = \varphi$ on γ.

Suppose that $A_i \in \partial Q$ or $A_i \notin \bar{Q}$, $i = 1, 2, \cdots, M$, then we define the spaces $H^{m,l}_\beta(Q)$, $l \geqq 0$ as in § 1.2. Let $H^{m-1/2,l-1/2}_\beta(\gamma)$, $m \geqq 1$, $l \geqq 0$ be the set of all functions φ on γ such that there exists $f \in H^{m,l}_\beta(Q)$ with $\varphi = f|_\gamma$ and

$$\|\varphi\|_{H^{m-1/2,l-1/2}_\beta(\gamma)} = \inf \|f\|_{H^{m,l}_\beta(Q)}.$$

where the infimum is taken over all functions $f \in H^{m,l}_\beta(\Omega)$ such that $f|_\gamma = \varphi$.

By $L_2(\gamma)$ we denote the space of the square integrable functions on γ. We also define the space $B^l_\beta(Q)$, $0 \leqq l \leqq 2$ analogously as in (1.2) replacing Ω by Q. Finally let $B^{l-1/2}_\beta(\gamma)$, $0 \leqq l \leqq 2$, be the space of all functions φ for which there exists $f \in B^l_\beta(Q)$ such that $f = \varphi$ on γ.

Remark 2. Although $B^l_\beta(Q)$, $0 \leqq l \leqq 1$ is not a subspace of $H^1(Q)$ the trace of $f \in B^l_\beta(Q)$ on γ obviously exists.

Remark 3. The norms $\|\cdot\|_{H^{m-1/2}(\gamma)}$ and $\|\cdot\|_{H^{m-1/2,l-1/2}_\beta(\gamma)}$ obviously depend on Q.

Remark 4. In what will follow Q will often be the polygonal domain and γ some of its edges. Although the set $H^{m-1/2,l-1/2}_\beta(\gamma)$ is characterized only by β_i associated to the vertices of the edges γ, we are defining the space $H^{m-1/2,l-1/2}_\beta(\gamma)$ depending on $\beta = (\beta_1, \cdots, \beta_M)$.

2. Regularity of the solution of the Poisson problem on a polygonal domain. In this chapter we will discuss the regularity of the problem

$$\begin{aligned} -\Delta u &= f \quad \text{on } \Omega, \\ u &= g^0 \quad \text{on } \Gamma^0, \\ \frac{\partial u}{\partial n} &= g^1 \quad \text{on } \Gamma^1 \end{aligned} \tag{2.1}$$

where

$$\Gamma^0 = \bigcup_{i \in \mathscr{D}} \bar{\Gamma}_i, \qquad \Gamma^1 = \Gamma - \Gamma^0.$$

Γ^0 will be called the Dirichlet boundary, Γ^1 the Neumann boundary. If $\Gamma^0 = \Gamma$ (respectively $\Gamma^1 = \Gamma$), then we will speak about the Dirichlet (respectively Neumann) problem. If $\Gamma^0 \neq \Gamma$ and $\Gamma^1 \neq \Gamma$, then we will speak about the mixed problem. The main theorem of this chapter is:

THEOREM 2.1. *Let* $f \in B^0_\beta(\Omega)$, $g^i \in B^{3/2-i}_\beta(\Gamma^i)$, $i=0,1$, $\beta=(\beta_1,\cdots,\beta_M)$, $0<\beta_i<1$, $\beta_i>1-\pi/\omega_i$ (*respectively* $\beta_i>1-\pi/2\omega_i$ *if Dirichlet and Neumann boundary conditions are imposed on the edges* Γ_{i-1}, Γ_i, $\bar\Gamma_{i-1}\cap\bar\Gamma_i=A_i$) *and let* $\Gamma^0\neq\varnothing$. *Then the problem* (2.1) *has a unique solution* u *in* $H^1(\Omega)$ *and* $u\in B^2_\beta(\Omega)$. □

Remark 1. If $\Gamma^0=\varnothing$ then the theorem still holds provided that f and g satisfy the condition (2.38) and the uniqueness is understood modulo a constant function.

Remark 2. g^1 should be understood as the vector $g^1=(g^1_1, g^1_2,\cdots, g^1_p)$; p is an integer $\leqq M$ such that $g^1_l=G^1_l|_{\Gamma_{i_l}}$, $\cup_{l=1}^p \Gamma_{i_l}=\Gamma^1$, $G^1_l\in B^1_\beta(\Omega)$, and $\|G^1\|^2_{H^{k,1}_\beta(\Omega)}=\sum_{l=1}^p \|G^1_l\|^2_{H^{k,1}_\beta(\Omega)}$.

Remark 3. It can be seen from the proof of the theorem that if $f\in H^{k,0}_\beta(\Omega)$, $G^j\in H^{k+2-j,2-j}_\beta(\Omega)$, $j=0,1$, $\beta_i>1-\pi/\omega_i$, (respectively $\beta_i>1-\pi/2\omega_i$) and $k\geqq 0$, then the solution of (2.1) exists in $H^{k+2,2}_\beta$ and

$$\|u\|_{H^{k+2,2}_\beta(\Omega)}\leqq C(k)\left(\|f\|_{H^{k,0}_\beta(\Omega)}+\sum_{j=0,1}\|G^j\|_{H^{k+2-j,2-j}_\beta(\Omega)}\right)$$

which is a kind of the "shift" theorem. Usually the shift theorem is expressed in the terms of usual Sobolev spaces so that

$$u=w+\sum_{i=1}^{m(k)} C_i\varphi_i$$

where φ_i are singular functions and for w there is the same shift theorem as for the domain with smooth boundary and without specific estimates of various constants in dependence on k. Theorem 2.1 is related to the known results but the authors were unable to find the theorem characterizing the solution in the framework of the countable normed space $B^2_\beta(\Omega)$ which is essential for applications.

Remark 4. The singular functions φ_i are associated to the vertices A_i, $\varphi_i=r^\alpha 1g^q r\psi(\theta)$, $\alpha>0$, $q\geqq 0$ integer, and $\psi(\theta)$ is an analytic function of θ. (r,θ) are the polar coordinates with the origin at A_i. Function φ_i belongs to the same space $B^2_\beta(\Omega)$, $\beta\in(1-\alpha,1)$ independent of q (see Remark 1 in § 1).

Remark 5. The proof of the theorem utilizes simple expansions of the solution, although this reasoning is very special. This approach is used to illuminate the main idea which will be used in the second paper of the series in an abstract form without using explicitly the mentioned expansion argument.

2.1. Auxiliary problems on the cone and the strip. Let

$$Q=S^\omega_\infty=\{r,\theta\,|\,0<r<\infty,\ 0<\theta<\omega\},$$
$$\Gamma_1=\{r,\theta\,|\,0<r<\infty,\ \theta=0\},$$
$$\Gamma_2=\{r,\theta\,|\,0<r<\infty,\ \theta=\omega\},$$

and

$$D=\{\tau,\theta\,|\,-\infty<\tau<\infty,\ 0<\theta<\omega\},$$
$$\tilde\Gamma_1=\{\tau,\theta\,|\,-\infty<\tau<\infty,\ \theta=0\},$$
$$\tilde\Gamma_2=\{\tau,\theta\,|\,-\infty<\tau<\infty,\ \theta=\omega\}.$$

The spaces $\mathscr{K}^{k,l}_\beta(Q)$, $0\leqq l\leqq 2$ and $\mathscr{K}^k_h(D)$, $k\geqq 0$, we defined in § 1.2. Let $C^\infty_\#(Q)$ be the collection of infinitely differentiable functions on $\bar Q$ such that:

for any $u\in C^\infty_\#(Q)$ there exists a positive number $A=A(u)$ such that u vanishes on $Q-Q_A$ where $Q_A=\{(r,\theta)\,|\,1/A<r<A,\ 0<\theta<\omega\}$.

Analogously we denote by $C^\infty_{\#}(D)$ the collection of infinitely differentiable functions on $\bar{D}$ such that for any $u \in C^\infty_{\#}(D)$ there exists $A = A(u) > 0$ such that u vanishes on $D - D_A$ where $D_A = \{(\tau, \theta) \mid -A < \tau < A,\ 0 < \theta < \omega\}$. It is not difficult to show (see [12]):

LEMMA 2.1. *$C^\infty_{\#}(Q)$ (respectively $C^\infty_{\#}(D)$) is dense in $\mathcal{H}^{k,l}_\beta(Q)$ (respectively $\mathcal{H}^k_h(D)$), $k \geqq l \geqq 0$.* □

LEMMA 2.2. *The space $\mathcal{H}^{k,l}_\beta(Q)$ and $\mathcal{H}^k_h(D)$ are complete.* □

Consider now the following problem on Q,

$$
\begin{aligned}
&-\Delta u = -\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right) = f \quad \text{on } Q, \\
&u|_{\theta=0} = g^0 = G^0|_{\theta=0}, \\
&\left.\frac{\partial u}{\partial n}\right|_{\theta=\omega} = g^1 = G^1|_{\theta=\omega}
\end{aligned}
\tag{2.2}
$$

where g^0 and g^1 are the traces of functions G^0 and G^1 defined on Q. Introducing new variable

$$\tau = \ln \frac{1}{r}$$

we transform the problem (2.2) into the problem on D

$$-\left(\frac{\partial^2 \tilde{u}}{\partial \tau^2} + \frac{\partial^2 \tilde{u}}{\partial \theta^2}\right) = \tilde{f}(\tau, \theta), \tag{2.3a}$$

$$
\begin{aligned}
&\tilde{u}|_{\theta=0} = \tilde{g}^0 = \tilde{G}^0|_{\theta=0}, \\
&\left.\frac{\partial \tilde{u}}{\partial \theta}\right|_{\theta=\omega} = \tilde{g}^1 = \tilde{G}^1|_{\theta=\omega}
\end{aligned}
\tag{2.3b}
$$

where

$$
\begin{aligned}
&\tilde{u}(\tau, \theta) = u(e^{-\tau}, \theta), \tilde{f}(\tau, \theta) = e^{-2\tau} f(e^{-\tau}, \theta), \\
&\tilde{G}^l(\tau, \theta) = e^{-l\tau} G^l(e^{-\tau}, \theta), \qquad l = 0, 1.
\end{aligned}
$$

LEMMA 2.3. *Let $f \in \mathcal{L}_h(D)$, $\tilde{G}^i \in \mathcal{H}^{2-i}_h(D)$, $i = 0, 1$, $0 < h < \pi/2\omega$, then the solution $\tilde{u}$ of (2.3) exists in $\mathcal{H}^2_h(D)$, is unique and for $0 \leqq |\alpha| \leqq 2$:*

$$
\begin{aligned}
\|D^\alpha \tilde{u}\|^2_{\mathcal{L}_h(D)} &= \int_D e^{2h\tau} |D^\alpha \tilde{u}|^2 \, d\tau \, d\theta \\
&\leqq C\left[\|\tilde{f}\|^2_{\mathcal{L}_h(D)} + \sum_{i=0}^{1} \|\tilde{G}^i\|^2_{\mathcal{H}^{2-i}_h(D)}\right],
\end{aligned}
\tag{2.4}
$$

where

$$D^\alpha \tilde{u} = \frac{\partial^{|\alpha|} \tilde{u}}{\partial \tau^{\alpha_1} \partial \theta^{\alpha_2}}$$

and C is independent of $\tilde{f}$ and $\tilde{G}^i$.

Proof. (1) Because of Lemmas 2.1 and 2.2, we may assume that $\tilde{f}, \tilde{G}^i \in C^\infty_{\#}(D)$. Denote by $\hat{f}(\lambda, \theta) = \mathcal{F}(\tilde{f}) = 1/\sqrt{2\pi} \int_{-\infty}^{+\infty} e^{-i\lambda\tau} f(\tau, \theta)\, d\tau$, $\hat{G}^i(\lambda, \theta) = \mathcal{F}(\tilde{G}^i)$ the Fourier transform (in τ) of $\tilde{f}$ and $\tilde{G}^i$.

Because $\tilde{f}, \tilde{G}^i \in C^\infty_{\#}(D)$ the Fourier transform for all λ. By the basic properties of the Fourier transform we get with $\lambda = \xi + ih$, $-\infty < \xi < \infty$:

$$
\begin{aligned}
&-\frac{\partial \hat{u}}{\partial \theta^2}(\lambda, \theta) + \lambda^2 \hat{u}(\lambda, \theta) = \hat{f}(\lambda, \theta) \quad \text{for } \theta \in I = (0, \omega), \\
&\hat{u}(\lambda, \theta)|_{\theta=0} = \hat{g}^0 = \hat{G}^0(\lambda, \theta)|_{\theta=0}, \\
&\frac{\partial \hat{u}}{\partial \theta}(\lambda, \theta)|_{\theta=0} = \hat{g}^1 = \hat{G}^1(\lambda, \theta)|_{\theta=\omega}.
\end{aligned}
\tag{2.5}
$$

The boundary value problem for the ordinary differential equation

$$-\hat{u}''+\lambda^2\hat{u}=0,$$

$$\hat{u}|_{\theta=0}=\frac{\partial\hat{u}}{\partial\theta}\bigg|_{\theta=\omega}=0$$

has eigenvalues $\lambda_k=i(\pi/\omega)(k-\frac{1}{2})$, $k=1, 2, \cdots$, and corresponding eigenfunctions $u_k=\sin(\pi/\omega)(k-\frac{1}{2})\theta$. Hence for $0<h<\pi/2\omega$ (2.5) always has a unique solution and by [3], [14] (formula (1.14))

$$\begin{aligned}\|\hat{u}\|^2_{H^2(I)}+|\lambda|^4\|\hat{u}\|^2_{L_2(I)}\leqq C[&\|\hat{f}\|^2_{L_2(I)}+\|\hat{G}^0\|^2_{H^2(I)}+\|\hat{G}^1\|^2_{H^1(I)}\\&+|\lambda|^3|\hat{G}^0(\lambda,0)|^2+|\lambda||\hat{G}^1(\lambda,\omega)|^2].\end{aligned}\tag{2.6}$$

It follows from the basic property of the Fourier transform that for any integer s, $s\leqq k$ and for any F in the set of admissible functions

$$\begin{aligned}\int_0^\omega\int_{-\infty}^{+\infty}\left|\frac{\partial^k F(\tau,\theta)}{\partial\tau^s\partial\theta^{k-s}}\right|^2 e^{2h\tau}\,d\tau\,d\theta&=\int_0^\omega\left(\int_{-\infty}^{\infty}\left|e^{h\tau}\frac{\partial^k F(\tau,\theta)}{\partial\tau^s\partial\theta^{k-s}}\right|^2 d\tau\right)d\theta\\&=\int_0^\omega\int_{-\infty+ih}^{\infty+ih}|\lambda|^{2s}\left|\frac{\partial^{k-s}\hat{F}(\lambda,\theta)}{\partial\theta^{k-s}}\right|^2 d\lambda\,d\theta.\end{aligned}\tag{2.7}$$

Hence for $\tilde{u}=\mathscr{F}^{-1}(\hat{u})$ we get

$$\int_D e^{2h\tau}\left|\frac{\partial^2\tilde{u}}{\partial\theta^2}\right|^2 d\tau\,d\theta=\int_{-\infty+ih}^{\infty+ih}\left\|\frac{\partial^2\hat{u}}{\partial\theta^2}\right\|^2_{L_2(I)}d\lambda,\tag{2.8}$$

$$\int_D e^{2h\tau}\left|\frac{\partial^2\tilde{u}}{\partial\tau^2}\right|^2 d\tau\,d\theta=\int_{-\infty+ih}^{\infty+ih}|\lambda|^4\|\hat{u}\|^2_{L_2(I)}\,d\lambda.\tag{2.9}$$

By the interpolation space theorem [8]

$$\begin{aligned}|\lambda|^2\left\|\frac{\partial\hat{u}}{\partial\theta}\right\|^2_{L_2(I)}&\leqq|\lambda|^2\|\hat{u}\|^2_{H^1(I)}\leqq C|\lambda|^2\|\hat{u}\|_{L_2(I)}\|\hat{u}\|_{H^2(I)}\\&\leqq C|\lambda|^2\|\hat{u}\|_{L_2(I)}\left(\|\hat{u}\|_{L_2(I)}+\left\|\frac{\partial\hat{u}}{\partial\theta}\right\|_{L_2(I)}+\left\|\frac{\partial^2\hat{u}}{\partial\theta^2}\right\|_{L_2(I)}\right)\\&\leqq C\left(\left(\left(1+\frac{C}{2}\right)|\lambda|^2+\frac{1}{2}|\lambda|^4\right)\|\hat{u}\|^2_{L_2(I)}+\frac{1}{2C}|\lambda|^2\left\|\frac{\partial\hat{u}}{\partial\theta}\right\|^2_{L_2(I)}+\frac{1}{2}\left\|\frac{\partial^2\hat{u}}{\partial\theta^2}\right\|^2_{L_2(I)}\right)\\&\leqq C_1\left(|\lambda|^4\|\hat{u}\|^2_{L_2(I)}+\left\|\frac{\partial^2\hat{u}}{\partial\theta^2}\right\|^2_{L_2(I)}\right)\end{aligned}$$

where C_1 depends on h but not on λ and $\hat{u}$. Hence

$$\begin{aligned}\int_D e^{2h\tau}\left|\frac{\partial^2\tilde{u}}{\partial\tau\partial\theta}\right|^2 d\tau\,d\theta&=\int_{-\infty+ih}^{\infty+ih}|\lambda|^2\left\|\frac{\partial\hat{u}}{\partial\theta}\right\|^2_{L_2(I)}d\lambda\\&\leqq C_1\int_{-\infty+ih}^{\infty+ih}\left(|\lambda|^4\|\hat{u}\|^2_{L_2(I)}+\left\|\frac{\partial^2\hat{u}}{\partial\theta^2}\right\|^2_{L_2(I)}\right)d\lambda.\end{aligned}\tag{2.10}$$

We have also

$$\int_{-\infty+ih}^{\infty+ih} (\|\hat{f}\|^2_{L_2(I)} + \|\hat{G}^0\|^2_{H^2(I)} + \|\hat{G}^1\|^2_{H^1(I)})\, d\lambda \tag{2.11}$$
$$= \int_D e^{2h\tau}\left(|\tilde{f}|^2 + \sum_{l=0}^{2}\left|\frac{\partial^l \tilde{G}^0}{\partial\theta^l}\right|^2 + \sum_{l=0}^{1}\left|\frac{\partial^l \tilde{G}^1}{\partial\theta^l}\right|^2\right) d\tau\, d\theta,$$

$$\int_{-\infty+ih}^{\infty+ih} |\lambda|^3 |\hat{G}^0(\lambda, 0)|^2\, d\lambda \leqq C\|e^{h\tau}\tilde{G}^0(\lambda, 0)\|^2_{H^{3/2}(\mathbf{R}_1)} \tag{2.12}$$
$$\leqq C\|e^{h\tau}\tilde{G}^0(\lambda, \theta)\|^2_{H^2(D)},$$

$$\int_{-\infty+ih}^{\infty+ih} |\lambda|\,|\hat{G}^1(\lambda, \omega)|^2\, d\lambda \leqq C\|e^{h\tau}\tilde{G}^1(\lambda, \omega)\|^2_{H^{1/2}(\mathbf{R}_1)} \tag{2.13}$$
$$\leqq C\|e^{h\tau}\tilde{G}^1(\lambda, \theta)\|^2_{H^1(D)}.$$

Hence from (2.6) using (2.8)-(2.13) we get for $|\alpha| = 2$;

$$\int_D e^{2h\tau}|D^\alpha \tilde{u}|^2\, d\tau\, d\theta \leqq C\left[\|\tilde{f}\|^2_{L_h(D)} + \sum_{i=0}^{1}\|\tilde{G}^i\|^2_{\mathscr{H}_h^{2-i}(D)}\right]. \tag{2.14}$$

For $l = 0, 1$ we have

$$\int_D e^{2h\tau}\left|\frac{\partial^l \tilde{u}}{\partial\tau^l}\right|^2 d\tau\, d\theta = \int_{-\infty+ih}^{\infty+ih} |\lambda|^{2l}\|\hat{u}\|^2_{L_2(I)}\, d\lambda$$
$$\leqq h^{-2l}\int_{-\infty+ih}^{\infty+ih} |\lambda|^4\|\hat{u}\|^2_{L_2(I)}\, d\lambda \tag{2.15}$$
$$\leqq Ch^{-2l}\left(\|\tilde{f}\|^2_{H_h(D)} + \sum_{i=0}^{1}\|\tilde{G}^i\|^2_{\mathscr{H}_h^{2-i}(D)}\right),$$

$$\int_D e^{2h\tau}\left|\frac{\partial\tilde{u}}{\partial\theta}\right|^2 d\tau\, d\theta = \int_{\infty-ih}^{\infty+ih} |\hat{u}|^2_{H^1(I)}\, d\lambda$$
$$\leqq h^{-2}\int_{-\infty+ih}^{\infty+ih} |\lambda|^2|\hat{u}|^2_{H^1(I)}\, d\lambda \tag{2.16}$$
$$\leqq Ch^{-2}\left(\|\tilde{f}\|^2_{\mathscr{L}_h(D)} + \sum_{i=0}^{1}\|\tilde{G}^i\|^2_{\mathscr{H}_h^{2-i}(D)}\right)$$

and (2.4) is proven.

(2) Let $\tilde{u} \in \mathscr{H}_h^2(D)$ and

$$\Delta\tilde{u} = 0 \quad \text{in } D,$$
$$\tilde{u}|_{\theta=0} = \left.\frac{\partial\tilde{u}}{\partial\theta}\right|_{\theta=\omega} = 0.$$

Hence we can write

$$\tilde{u}(\tau, \theta) = \sum_{j=1}^{\infty} a_j(\tau)\sin\frac{\pi\theta}{\omega}\left(j - \frac{1}{2}\right), \tag{2.17}$$

$$a_j(\tau) = \frac{2}{\omega}\int_0^\omega \tilde{u}(\tau, \theta)\sin\frac{\pi\theta}{\omega}\left(j - \frac{1}{2}\right) d\theta, \tag{2.18}$$

$$a_j''(\tau)=\frac{2}{\omega}\int_0^{\omega}\frac{\partial^2\tilde{u}(\tau,\theta)}{\partial\tau^2}\sin\frac{\pi\theta}{\omega}\left(j-\frac{1}{2}\right)d\theta$$

$$=\left(\frac{\pi}{\omega}\left(j-\frac{1}{2}\right)\right)^2 a_j(\tau).$$

Because for each j, $j=1, 2, \cdots$, $a_j(\tau)$ satisfies

$$a_j''(\tau)-\left(\frac{\pi}{\omega}\left(j-\frac{1}{2}\right)\right)^2 a_j(\tau)=0$$

we have

$$a_j(\tau)=c_j\,e^{-(\pi/\omega)(j-1/2)\tau}+d_j\,e^{(\pi/\omega)(j-1/2)\tau}. \tag{2.19}$$

For $A>0$ arbitrary

$$\infty>\int_{D_A}e^{2h\tau}|\tilde{u}|^2\,d\tau\,d\theta$$

$$=\int_{-A}^{A}\left(\int_0^{\omega}\left(\sum_{j=1}^{\infty}a_j(\tau)\sin\frac{\pi\theta}{\omega}\left(j-\frac{1}{2}\right)\right)^2 d\theta\right)e^{2h\tau}\,d\tau$$

$$=\int_{-A}^{+A}\frac{\omega}{2}\sum_{j=1}^{\infty}|a_j(\tau)|^2\,e^{2h\tau}\,d\tau$$

$$=\frac{\omega}{2}\sum_{j=1}^{\infty}\int_{-A}^{+A}|a_j(\tau)|^2\,e^{2h\tau}\,d\tau.$$

For $j=1$

$$|a_1(\tau)|^2=c_1^2\,e^{-(\pi/\omega)\tau}+d_1^2\,e^{(\pi/\omega)\tau}+2c_1d_1$$

and hence

$$\int_{-A}^{A}|a_1(\tau)|^2\,e^{2h\tau}\,d\tau=2c_1d_1\frac{1}{2h}(e^{2hA}-e^{-2hA})+\frac{c_1^2}{(2h-\pi/\omega)}(e^{(2h-\pi/\omega)A}-e^{-(2h-\pi/\omega)A})$$

$$+\frac{d_1^2}{(2h+\pi/\omega)}(e^{(2h+\pi/\omega)A}-e^{-(2h+\pi/\omega)A})$$

and

$$\infty>\lim_{A\to\infty}\int_{-A}^{+A}|a_1(\tau)|^2\,e^{2h\tau}\,d\tau$$

$$=\lim_{A\to\infty}\left(\frac{c_1d_1}{h}\,e^{2hA}+\frac{c_1^2}{(\pi/\omega-2h)}\,e^{|(\pi/\omega-2h)|A}+\frac{d_1^2}{(2h+\pi/\omega)}\,e^{(2h+\pi/\omega)A}\right).$$

Hence c_1, $d_1=0$. Similarly we get $c_j=d_j=0$ for all $1\leqq j<\infty$, which implies the uniqueness of the solution of (2.3) in $\mathcal{H}_h^2(D)$. □

We have not explicitly used in the proof of Lemma 2.3 the assumption $h<\pi/2\omega$. We have used only $h>0$ and $h\neq\pi/2\omega(k-\frac{1}{2})$, $k=1, 2, \cdots$. The assumption $h<\pi/2\omega$ is essential in the next lemma.

LEMMA 2.4. *Let the assumptions of Lemma* 2.3 *hold. Let, in addition,* $\tilde{f}(\tau, \theta) = 0$, $\tilde{G}^i(\tau, \theta) = 0$ *for* $\tau < 0$. *Then for* $\varepsilon > 0$ *and* $0 \leqq \gamma = h + (\pi/2\omega) - \varepsilon$, $\tilde{D} = \{\tau, \theta \,|\, -\infty < \tau < 0, 0 < \theta < \omega\}$ *one has*

$$\int_{\tilde{D}} |D^\alpha \tilde{u}|^2 \, e^{2(h-\gamma)\tau} \, d\tau \, d\theta \leqq C(\varepsilon) \int_{\tilde{D}} |D^\alpha \tilde{u}|^2 \, e^{2h\tau} \, d\tau \, d\theta, \qquad |\alpha| \leqq 2. \tag{2.20}$$

Proof. Equations (2.17)-(2.19), hold, and

$$\infty > \int_{\tilde{D}} |\tilde{u}|^2 \, e^{2h\tau} \, d\tau \, d\theta = \frac{\omega}{2} \int_{-\infty}^{0} \sum_{j=1}^{\infty} |a_j(\tau)|^2 \, e^{2h\tau} \, d\tau.$$

Hence for $A > 0$ arbitrary

$$\int_{-A}^{0} |a_j(\tau)|^2 \, e^{2h\tau} \, d\tau = 2c_j d_j (1 - e^{-2hA}) + \frac{c_j^2}{2(h - (\pi/\omega)(j - \frac{1}{2}))} (1 - e^{-2(h-(\pi/\omega)(j-1/2))A})$$
$$+ \frac{d_j^2}{2(h + (\pi/\omega)(j - \frac{1}{2}))} (1 - e^{-2(h+(\pi/\omega)(j-1/2))A}).$$

Since $0 < h < \pi/2\omega$ and $A > 0$ is arbitrary we have $c_j = 0$, $j = 1, 2, \cdots$ and

$$\int_{-A}^{0} |a_j(\tau)|^2 \, e^{2h\tau} \, d\tau = \frac{d_j^2}{2(h + (\pi/\omega)(j - \frac{1}{2}))} (1 - e^{-2(h+(\pi/\omega)(j-1/2))A}),$$

$$\int_{-A}^{0} |a_j(\tau)|^2 \, e^{2(h-\gamma)\tau} \, d\tau = \frac{d_j^2}{2(h - \gamma + (\pi/\omega)(j - \frac{1}{2}))} (1 - e^{-2(h-\gamma+(\pi/\omega)(j-\frac{1}{2}))A}).$$

If $0 \leqq \gamma = h + \pi/2\omega - \varepsilon$, $\varepsilon > 0$ then

$$\int_{\tilde{D}} |\tilde{u}(\tau, \theta)|^2 \, e^{2(h-\gamma)\tau} \, d\tau \, d\theta = \int_{\tilde{D}} |\tilde{u}(\tau, \theta)|^2 \, e^{2(-\pi/2\omega+\varepsilon)\tau} \, d\tau \, d\theta$$
$$= \frac{\omega}{2} \lim_{A \to \infty} \sum_{j=1}^{\infty} \frac{d_j^2 (1 - e^{-2((\pi/\omega)(j-1)+\varepsilon)A})}{2(\varepsilon + (\pi/\omega)(j-1))}$$
$$\leqq C \frac{\omega}{2} \sum_{j=1}^{\infty} \frac{d_j^2}{2(h + (\pi/\omega)(j - \frac{1}{2}))}$$
$$= C \int_{\tilde{D}} |\tilde{u}|^2 \, e^{2h\tau} \, d\tau \, d\theta.$$

Similarly we have for $|\alpha| \leqq 2$

$$\int_{\tilde{D}} |D^\alpha \tilde{u}|^2 \, e^{2(h-\gamma)\tau} \, d\tau \, d\theta \leqq C \int_{\tilde{D}} |D^\alpha \tilde{u}|^2 \, e^{2h\tau} \, d\tau \, d\theta. \qquad \square$$

Lemmas 2.3 and 2.4 address the regularity of the problem (2.3) when on Γ_1, respectively, Γ_2, the Dirichlet respectively the Neumann condition has been given. The same statement holds if on Γ_1 and Γ_2 the Dirichlet or Neumann condition are given.

LEMMA 2.5. *Let* $\tilde{f} \in \mathcal{H}_h(D)$, $\tilde{G}^0 \in \mathcal{H}_h^2(D)$ *(respectively* $\tilde{G}^1 \in \mathcal{H}_h^1(D)$*),* $0 < h < \pi/\omega$. *Then the Dirichlet (respectively Neumann) problem*

$$-\left(\frac{\partial^2 \tilde{u}}{\partial \tau^2} + \frac{\partial^2 u}{\partial \theta^2}\right) = \tilde{f}(\tau, \theta), \tag{2.21a}$$

$$\tilde{u}\Big|_{\substack{\theta=0 \\ \theta=\omega}} = \tilde{g}^0 = \tilde{G}^0\Big|_{\substack{\theta=0 \\ \theta=\omega}} \tag{2.21b}$$

(*respectively*)

$$\left.\frac{\partial \tilde{u}}{\partial n}\right|_{\substack{\theta=0\\ \theta=\omega}} = \tilde{g}^1 = \tilde{G}^2\Big|_{\substack{\theta=0\\ \theta=\omega}}, \tag{2.21c}$$

has a unique solution in $\mathscr{H}^2_h(D)$ *and for* $0 \leqq |\alpha| \leqq 2$

$$\|D^\alpha \tilde{u}\|^2_{\mathscr{L}_h(D)} \leqq C[\|\tilde{f}\|^2_{\mathscr{L}_h(D)} + \|\tilde{G}^0\|^2_{\mathscr{H}^2_h(D)} \quad (\textit{respectively } \|\tilde{G}^1\|^2_{\mathscr{H}^1_h(D)})]. \tag{2.22}$$

If in addition $\tilde{f} = 0$, $\tilde{G}^0$ (*respectively* $\tilde{G}^1$) $= 0$ *for* $\tau < 0$ *then for* $0 \leqq \gamma = h + \pi/\omega - \varepsilon$, $\varepsilon > 0$, $0 \leqq |\alpha| \leqq 2$

$$\int_{\tilde{D}} |D^\alpha \tilde{u}^*|^2 \, e^{2(h-\gamma)\tau} \, d\tau \, d\theta \leqq C \int_{\tilde{D}} |D^\alpha \tilde{u}|^2 \, e^{2h\tau} \, d\tau \, d\theta \tag{2.23}$$

where $\tilde{u}^*(\tau, \theta) = \tilde{u}(\tau, \theta)$, *for the Dirichlet problem and*

$$\tilde{u}^*(\tau, \theta) = \tilde{u}(\tau, \theta) - \frac{1}{\omega} \int_0^\omega \tilde{u}(\tau, \theta) \, d\theta$$

for the Neumann problem. □

The proof is quite the same. In the case of Neumann conditions it is enough to realize that a similar summation as in (2.17) is for $j = 0, 1, 2, \cdots$.

LEMMA 2.6. *Let* $f \in \mathscr{L}_\beta(Q)$, $g^i \in \mathscr{H}^{2-i,2-i}_\beta(Q)$, $i = 0, 1$, $0 < \beta < 1$, $\beta > 1 - \pi/2\omega$ *and let* $f = G^i = 0$ *for* $r \geqq 1$; *then the mixed problem*

$$-\Delta u = f(r, \theta), \tag{2.24a}$$

$$u|_{\theta=0} = g^0 = G^0|_{\theta=0}, \qquad \left.\frac{\partial u}{\partial n}\right|_{\theta=\omega} = \frac{1}{r}\left.\frac{\partial u}{\partial \theta}\right|_{\theta=\omega} = g^1 = G^1|_{\theta=\omega} \tag{2.24b}$$

has a solution such that

(i) $(u - G^0) \in W^2_\beta(Q)$, $u \in \mathscr{H}^{2,2}_\beta(S_1)$,

(ii) $\|\mathscr{D}^1 u\|_{H^0(Q)} < \infty$,

(iii) *there exists a constant* C *independent of* u, f, G^i *such that for* $|\alpha| \leqq 2$

$$\|u\|^2_{\mathscr{H}^{2,2}_\beta(S_1)} \leqq C\left[\|f\|^2_{\mathscr{L}_\beta(Q)} + \sum_{i=0}^{1} \|G^i\|^2_{\mathscr{H}^{2-i,2-i}_\beta(Q)}\right] \tag{2.25}$$

where $Q = \{r, \theta \,|\, 0 < r < \infty, 0 < \theta < \omega\}$ *and* $S_1 = \{r, \theta \,|\, 0 < r < 1, 0 < \theta < \omega\}$.

Proof. First assume that $G^0 = 0$. Let $0 < h = 1 - \beta < \pi/2\omega$ and $\tau = \ln(1/r)$. Then we have for $\tilde{f}(\tau, \theta) = e^{-2\tau} f(e^{-\tau}, \theta)$ and $\tilde{G}^1(\tau, \theta) = e^{-\tau} G^1(e^{-\tau}, \theta)$

$$\begin{aligned} \int_D e^{2h\tau} |\tilde{f}(\tau, \theta)|^2 \, d\tau \, d\theta &= \int_D e^{-2(2-h)\tau} |f(e^{-\tau}, \theta)|^2 \, d\tau \, d\theta \\ &= \int_Q r^{2\beta} |f(r, \theta)|^2 r \, dr \, d\theta \\ &= \|f\|^2_{\mathscr{L}_\beta(Q)}, \end{aligned} \tag{2.26a}$$

$$\begin{aligned} \int_D e^{2h\tau} |\tilde{G}^1(\tau, \theta)|^2 \, d\tau \, d\theta &= \int_Q r^{2(\beta-1)} |G^1(r, \theta)|^2 r \, dr \, d\theta \\ &= \int_{S_1 \cup Q_2} r^{2(\beta-1)} |G^1(r, \theta)|^2 r \, dr \, d\theta \end{aligned} \tag{2.26b}$$

where we denoted

$$Q_2 = \{(r, \theta) \mid 1 \leqq r < \infty, 0 < \theta < \omega\}.$$

By Lemma A.2 (see Appendix)

$$\int_{S_1} r^{2(\beta-1)} |G^1|^2 r\, dr\, d\theta \leqq C\left[\sum_{|\alpha|=1} \int_{S_1} r^{2(\beta+\alpha_1-1)} |\mathscr{D}^\alpha G^1|^2 r\, dr\, d\theta + \int_{S_1} |G^1|^2 r\, dr\, d\theta\right]$$

and

$$\int_{Q_2} r^{2(\beta-1)} |G^1|^2 r\, dr\, d\theta \leqq \int_{Q_2} |G^1|^2 r\, dr\, d\theta.$$

Hence

$$\|\tilde{G}^1\|_{\mathscr{L}_h(D)} \leqq C \|G^1\|_{H^{1,1}_\beta(Q)}$$

and for $\alpha_1 + \alpha_2 = 1$

$$\int_D e^{2h\tau} |\tilde{G}^1_{\tau^{\alpha_1}\theta^{\alpha_2}}|^2\, d\tau\, d\theta = \int_Q r^{2(\alpha_1-1+\beta)} |G^1_{r^{\alpha_1}\theta^{\alpha_2}}|^2 r\, dr\, d\theta.$$

Therefore

$$\|\tilde{G}^1\|_{\mathscr{H}^1_h(D)} \leqq C \|G^1\|_{\mathscr{H}^{1,1}_\beta(Q)}. \tag{2.27}$$

Using (2.26) and (2.27) and Lemma 2.3 we see that the equation

$$\begin{aligned} &-(\tilde{u}_{\tau\tau} + \tilde{u}_{\theta\theta}) = \tilde{f} \quad \text{in } D, \\ &\tilde{u}|_{\theta=0} = 0, \\ &\left.\frac{\partial \tilde{u}}{\partial \theta}\right|_{\theta=\omega} = \tilde{G}^1|_{\theta=\omega} \end{aligned} \tag{2.28}$$

has the unique solution $\tilde{u} \in \mathscr{H}^2_h(D)$ and (2.4) holds. Let $u = \tilde{u}(\ln(1/r), \theta)$, then u satisfies (2.24) and for $|\alpha| \leqq 2$

$$\begin{aligned} \int_Q r^{2(\alpha_1-2+\beta)} |u_{r^{\alpha_1}\theta^{\alpha_2}}|^2 r\, dr\, d\theta &\leqq C \sum_{|\alpha|\leqq 2} \int_D e^{2h\tau} |\tilde{u}_{\tau^{\alpha_1}\theta^{\alpha_2}}|^2\, d\tau\, d\theta \\ &\leqq C[\|\tilde{f}\|^2_{\mathscr{L}_h(D)} + \|\tilde{G}^1\|^2_{\mathscr{H}^1_h(D)}] \\ &= C[\|f\|^2_{\mathscr{L}_\beta(Q)} + \|G^1\|^2_{\mathscr{H}^{1,1}_\beta(Q)}]. \end{aligned} \tag{2.29}$$

Hence $u \in W^2_\beta(Q)$ and (2.25) is proven for $G^0 = 0$. For $G^0 \neq 0$ we define $w = u - G^0$; then

$$\begin{aligned} &-\Delta w = f + \Delta G^0 = \bar{f}, \\ &w|_{\theta=\omega} = 0, \\ &\frac{1}{r}\left.\frac{\partial w}{\partial \theta}\right|_{\theta=\omega} = \left(G^1 - \frac{1}{r}\frac{\partial G^0}{\partial \theta}\right)\Bigg|_{\Gamma_2} = \bar{G}^1|_{\Gamma_2}. \end{aligned}$$

Applying now (2.29) (respectively (2.25)) to this case we get $w \in W^2_\beta(Q)$ and

$$\begin{aligned} \|w\|^2_{\mathscr{H}^{2,2}_\beta(S_1)} &\leqq C[\|\bar{f}\|^2_{\mathscr{L}_\beta(Q)} + \|\bar{G}^1\|^2_{\mathscr{H}^{1,1}_\beta(Q)}] \\ &\leqq C\left[\|f\|^2_{\mathscr{L}_\beta(Q)} + \sum_{i=0,1} \|G^i\|^2_{\mathscr{H}^{2-i,2-i}_\beta(Q)}\right] \end{aligned}$$

which proves (2.25) in full generality.

Let us prove now that $\|\mathscr{D}^1 u\|_{H^0(Q)} < \infty$. Equation (2.25) shows that $\|\mathscr{D}^1 u\|_{H^0(S_1)} < \infty$; hence we have to prove only that

$$\|\mathscr{D}^1 u\|_{H^0(Q_2)}, \qquad Q_2 = Q - \bar{S}_1.$$

We have by Lemma 2.4 for $h = 1-\beta$ and $0 \leqq \gamma = \pi/2\omega + h - \varepsilon$, $\varepsilon > 0$, $|\alpha| \leqq 2$

$$\int_{Q_2} |\mathscr{D}^\alpha u|^2 r^{2(\alpha_1 - 2 + \beta) + 2\gamma} r\, dr\, d\theta = \int_{\tilde{D}} |D^\alpha \tilde{u}|^2 \, e^{2(h-\gamma)\tau}\, d\tau\, d\theta < \infty. \tag{2.30}$$

Particularly for $\alpha_1 + \alpha_2 = 1$ and $0 < \varepsilon < \pi/2\omega$ we have

$$\begin{aligned} \int_{Q_2} |u_r|^2 r\, dr\, d\theta &\leqq \int_{Q_2} |u_r|^2 r^{2(-1+\beta+1-\beta+(\pi/2\omega)-\varepsilon)} r\, dr\, d\theta \\ &\leqq \int_{Q_2} |u_r|^2 r^{2(\alpha_1 - 2 + \beta) + 2\gamma} r\, dr\, d\theta < \infty \end{aligned} \tag{2.31}$$

and

$$\begin{aligned} \int_{Q_2} \frac{1}{r^2} |u_\theta|^2 r\, dr\, d\theta &\leqq \int_{Q_2} |u_\theta|^2 r^{2(-2+\beta+(1-\beta)+(\pi/2\omega)-\varepsilon)} \\ &\leqq \int_{Q_2} |u_\theta|^2 r^{2(\alpha_1 - 2 + \beta) + 2\gamma} r\, dr\, d\theta < \infty \end{aligned} \tag{2.32}$$

and (ii) of Lemma 2.6 is proven. □

Analogously we have the following.

LEMMA 2.7. *Let* $f \in \mathscr{L}_\beta(Q)$, $G^i \in \mathscr{H}_\beta^{2-i,2-i}(Q)$, $i = 0,\ 1$, $0 < \beta < 1$, $\beta > 1 - (\pi/\omega)$ $f = G^i = 0$ *for* $r > 1$, *then the Dirichlet (respectively Neumann) problem*

$$-\Delta u = f(r, \theta), \tag{2.33a}$$

$$u|_{\theta = 0,\omega} = g^0 = G^0|_{\theta = 0,\omega}$$
$$\left(\textit{respectively} \left.\frac{\partial u}{\partial n}\right|_{\theta = 0,\omega} = g^1 = G^1|_{\theta = 0,\omega} \right) \tag{2.33b}$$

has a solution such that

(i) $(u - G^0) \in W_\beta^2(Q)$ *(respectively* $u \in W_\beta^2(Q)$*)*, $u \in \mathscr{H}_\beta^{2,2}(S_1)$, *and*

(ii) $\|\mathscr{D}^1 u^*\|_{H^0(Q)} < \infty$

where $u^*(r, \theta) = u(r, \theta) - 1/\omega \int_0^\omega u(r, \theta)\, d\theta$ *for the Neumann problem,*

(iii) (2.25) *holds with* $G^1 = 0$ *(respectively* $G^0 = 0$*).*

Let us now prove the following.

LEMMA 2.8. *Let* $u \in \tilde{H}^1(Q) = \{u \mid \int_Q |\mathscr{D}^1 u|^2 r\, dr\, d\theta < \infty,\ u|_{\Gamma^0} = 0\}$ *and* $u = 0$ *for* $r > 1$ *be the solution of the problem*

$$-\Delta u = f,$$
$$u|_{\Gamma^0} = G^0|_{\Gamma^0},$$
$$\left.\frac{\partial u}{\partial n}\right|_{\Gamma^1} = G^1|_{\Gamma^1}$$

with $f \in \mathscr{L}_\beta(Q)$, $G^i \in \mathscr{H}_\beta^{2-i,2-i}(Q)$, f *and* G^i *vanish for* $r > 1$, $i = 0,\ 1$, $0 < \beta < 1$, $\beta > 1 - \pi/2\omega$ *for the mixed problem and* $\beta > 1 - \pi/\omega$ *for the Dirichlet and Neumann problem, where* Γ^0 *is the union of the edges of* Q, *or the empty set,* $\Gamma^1 = \partial Q - \Gamma^0$.

Further let $w \in W^2_\beta(Q)$ *be the solution of the same problem given in Lemmas* 2.6 *and* 2.7. *Then* $u = w$ *if* $\Gamma^0 \neq \emptyset$ (*i.e. for the Dirichlet or mixed problem*) *and* $u = w + C$ *if* $\Gamma^0 = \emptyset$ (*i.e. Neumann problem*).

Proof. We first prove the lemma for the Dirichlet problem. We may assume $G^0 = 0$. Since $\|\mathscr{D}^1 w\|_{H^0(Q)} < \infty$ by Lemma 2.7, we have for every $v \in \tilde{H}^1_0(Q) = \{v | \int_Q |\mathscr{D}^1 v|^2 r\, dr\, d\theta < \infty,\ v|_{\Gamma^0} = 0,\ v = 0 \text{ for } r > A(v)\}$

$$\int_Q \left(w_r v_r + \frac{1}{r^2} w_\theta v_\theta \right) r\, dr\, d\theta = \int_Q \left(u_r v_r + \frac{1}{r^2} u_\theta v_\theta \right) r\, dr\, d\theta$$

and hence

$$\int_Q \left((w_r - u_r) v_r + \frac{1}{r^2} (w_\theta - u_\theta) v_\theta \right) r\, dr\, d\theta = 0.$$

Because $\tilde{H}^1_0(Q)$ is dense in $\tilde{H}^1(Q)$ the equality holds also for $v = w - u$. This immediately gives $w = u + C$ and obviously $C = 0$. Now we prove the lemma for the Neumann problem.

Let $u^* = u - 1/\omega \int_0^\omega u(r, \theta)\, d\theta = u - b_0(r)$ and $w^* = w - \int_0^\omega (1/\omega) w(r, \theta)\, d\theta = w - a_0(r)$. Then by Lemma 2.7 $\|\mathscr{D}^1 w^*\|_{H^0(Q)} < \infty$. Let

$$\tilde{\tilde{H}}^1_0(Q) = \left\{ u \,\middle|\, \int_Q |\mathscr{D}^1 u|^2 r\, dr\, d\theta < \infty, \int_0^\omega u(r, \theta)\, d\theta = 0 \right\}.$$

Then for any $v \in \tilde{\tilde{H}}^1_0(Q)$ having bounded support

$$\begin{aligned} \int_Q \left(u_r v_r + \frac{1}{r^2} u_\theta v_\theta \right) r\, dr\, d\theta &= \int_Q \left(u^*_r v_r + \frac{1}{r^2} u^*_\theta v_\theta \right) r\, dr\, d\theta \\ &= \int_Q \left(w_r v_r + \frac{1}{r^2} w_\theta v_\theta \right) r\, dr\, d\theta \\ &= \int_Q \left(w^*_r v_r + \frac{1}{r^2} w^*_\theta v_\theta \right) r\, dr\, d\theta \end{aligned}$$

and hence

$$\int_Q \left((u^*_r - w^*_r) v_r + \frac{1}{r^2} (u^*_\theta - w^*_\theta) v_\theta \right) r\, dr\, d\theta = 0.$$

Since the set of $v \in \tilde{\tilde{H}}^1_0(Q)$ having bounded support is dense in $\tilde{\tilde{H}}^1_0(Q)$ we get $u^* - w^* = C$. Thus

$$u - w = u^* - w^* = a_0(r) - b_0(r) = C(r)$$

and because u, w solve the same problem and obviously $\partial C/\partial\theta = 0$, we get

$$\int_Q C_r v_r r\, dr\, d\theta = 0$$

for any $v \in H = \{v | \int_Q |\mathscr{D}^1 v|^2 r\, dr\, d\theta < \infty$, hence $C(r) = C_1 + C_2 \log(1/r)$; but $C(r) = (u - w) \in H^1(S_1)$ by Lemma 2.7, and hence $C_2 = 0$ and $u - w = C_1$. □

2.2. The regularity of the solution on a polygonal domain Ω.

LEMMA 2.9. *Let* $g = G|_\gamma \in H^{1/2,1/2}_\beta(\gamma)$ *where* γ *is the edge of* $\partial\Omega$. *Then* $(\Phi_{\beta/2} G)|_\gamma \in L_2(\gamma)$.

Proof. Let $\Gamma_i = \gamma$ be the edge connecting the vertices A_{i+1} and A_i, let A_i be placed at the origin and Γ_i lies on the x_1-axis. Assume that $S_\delta^{\omega_i} \subset \Omega$. It is sufficient to prove that on $\int_0^\delta r^\beta |G(x_1, 0)|^2 \, dx_1 < \infty$ with $\beta = \beta_i$.

Let $F = r^\beta G$. Then

$$F_{x_i} = r^\beta G_{x_i} + \beta r^{\beta-1} \frac{x_i}{r} G.$$

By Lemma A.3 of the Appendix

$$\|F_{x_i}\|_{\mathscr{L}_\beta(S_\delta^{\omega_i})} \leqq C \|G\|_{H_\beta^{1,1}(\Omega)}$$

and hence

$$\|F\|_{H^1(S_\delta^{\omega_i})} \leqq C \|G\|_{H_\beta^{1,1}(\Omega)}.$$

By the imbedding theorem $F \in L_p(I_\delta)$ for any $p > 1$, $I_\delta = (0, \delta)$ (see [1]), and

$$\|F\|_{L_p(I_\delta)} \leqq C \|F\|_{H^1(S_\delta^{\omega_i})} \leqq C \|G\|_\beta^{1,1}(\Omega).$$

Hence

$$\begin{aligned} \|r^{\beta/2} G\|_{L_2(I_\delta)}^2 &= \int_{I_\delta} r^{-\beta} |F|^2 \, dx_1 \\ &\leqq C \left(\int_{I_\delta} r^{-\beta q} \, dx_1 \right)^{1/q} \left(\int_{I_\delta} |F|^{2p} \right)^{1/p} \\ &\leqq C \|F\|_{L_{2p}(I_\delta)}^2 \leqq C \|F\|_{H^1(S_\delta^{\omega_i})}^2 \leqq C \|G\|_{H_\beta^{1,1}(\Omega)}^2 \end{aligned}$$

where $1/p + 1/q = 1$ and $\beta q < 1$, $p > 1$, $q > 1$. □

LEMMA 2.10. *Let $f \in \mathscr{L}_\beta(\Omega)$. Then $\int_\Omega fv \, dx$ is a linear continuous functional on $H^1(\Omega)$ and $\|f\|_{(H^1(\Omega))'} \leqq C \|f\|_{\mathscr{L}_\beta(\Omega)}$.*

The proof follows easily from the Schwarz inequality and imbedding theorem. See also [12]. □

LEMMA 2.11. *Let $f \in \mathscr{L}_\beta(\Omega)$, $G^i \in H_\beta^{2-i,2-i}(\Omega)$ $i = 0, 1$ and $|\Gamma^0| \neq 0$. Then*

$$\begin{aligned} &-\Delta u = f, \\ &u|_{\Gamma^0} = g^0 = G^0|_{\Gamma^0}, \\ &\left.\frac{\partial u}{\partial n}\right|_{\Gamma^1} = g^1 = G^1|_{\Gamma^1} \end{aligned} \tag{2.34}$$

has the unique solution $u \in H^1(\Omega)$ (in the weak sense) and

$$\|u\|_{H^1(\Omega)} \leqq C \left[\|f\|_{\mathscr{L}_\beta(\Omega)} + \sum_{i=0,1} \|G^i\|_{H_\beta^{2-i,2-i}(\Omega)} \right]. \tag{2.35}$$

Proof. Without loss of generality we can assume that $G^0 = 0$ because $\Delta G^0 \in \mathscr{L}_\beta(\Omega)$ and $\partial G^0 / \partial x_i \in H_\beta^{1,1}(\Omega)$. Applying Lemma 2.10 it suffices to show that

$$\int_{\Gamma^1} g^1 v \, dx$$

is a linear functional on $H^1(\Omega)$.

We have

$$\int_{\Gamma^1} g^1 v \, ds = \int_{\Gamma^1} r^{\beta/2} g^1 r^{-\beta/2} v \, ds, \tag{2.36}$$

$$\int_{\Gamma^1} r^{-\beta} v^2 \, ds \leqq \left(\int_{\Gamma^1} r^{-\beta p} \, ds \right)^{1/p} \left(\int_{\Gamma^1} |v|^{2q} \, dx \right)^{1/q} \leqq C \|v\|_{H^1(\Omega)}^2 \tag{2.37}$$

and (2.36) together with Lemma 2.9 and (2.37) shows that

$$\left| \int_{\Gamma^1} g^1 v \, ds \right| \leqq C \| G^1 \|_{H_\beta^{1,1}(\Omega)} \| v \|_{H^1(\Omega)}.$$

The Lax-Milgram lemma yields (2.35) and the uniqueness. □

Remark. If $|\Gamma^0| = 0$ and

$$\int_\Omega f \, dx + \int_{\Gamma^1} g^1 \, ds = 0, \tag{2.38}$$

then Lemma 2.11 holds in the factor space modulo a constant.

Proof of Theorem 2.1. Consider the polygonal domain shown in Fig. 2.1. Let

$$S_{i,\delta_i} = \{ r_i, \theta_i \,|\, 0 < r_i < \delta_i, 0 < \theta_i < \omega_i \} \subset \Omega$$

where (r_i, θ_i) are the polar coordinates with respect to the vertex A_i. (See Fig. 2.2.)

Let $\delta_i < 1$, be such that $S_{i,2\delta_i} \cap S_{j,2\delta_j} = \varnothing$ for $i \neq j$, $i, j = 1, \cdots M$. By Lemma 2.11 there is a unique solution of (2.1) in $H^1(\Omega)$. By Theorem 5.7.1, 5.7.1′ and 6.6.1 of [16], u is analytic in Ω and on Γ_i, $1 \leqq i \leqq M$ (because f and g^i are analytic functions in Ω

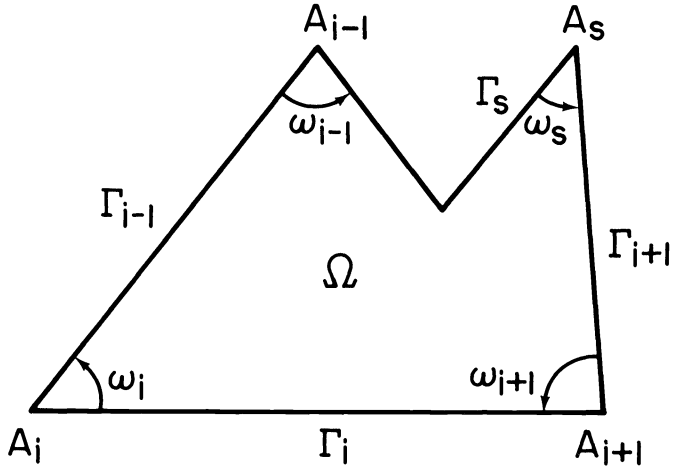


FIG. 2.1. *The polygonal domain.*

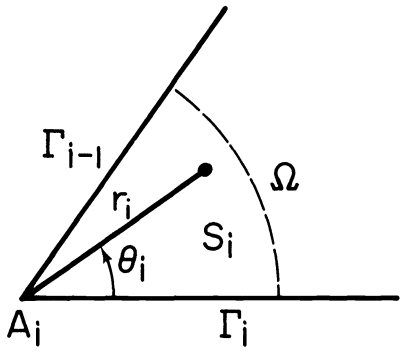


FIG. 2.2. *The scheme of coordinates* (r_i, θ_i).

and on Γ_i by our assumption). Hence Theorem 2.1 holds on $\Omega - \bigcup_{i=1}^{M} S_{i,\delta_i/4}$, and in particular we have for $|\alpha| = k$, $k \geqq 2$

$$\|r_i^{\alpha_1} \mathscr{D}^\alpha u\|_{\mathscr{L}_{\beta_i}(S_{i,\delta_i} - S_{i,\delta_i/2})} \leqq C_{i,0} d_{i,0}^{k-2} (k-2)!. \tag{2.39}$$

Hence, it is sufficient to prove that in each sector $S_{i,\delta_i/2}$, $1 \leqq i \leqq M$ and $|\alpha| = k$, $k \geqq 2$ we have

$$\|r_i^{\alpha_1 - 2} \mathscr{D}^\alpha u\|_{L_{\beta_i}(S_{i,\delta i/2})} \leqq L_i D_i^{k-2} P_i^{\alpha_2} (k-2)! \tag{2.40}$$

with L_i, D_i, P_i independent of k (see also Theorem 1.1 and (1.13)). There are three cases to be considered:

(i) $\Gamma_1 \subset \Gamma^0$, $\Gamma_{i-1} \subset \Gamma^1$,
(ii) Γ_i, $\Gamma_{i-1} \subset \Gamma^0$,
(iii) Γ_i, $\Gamma_{i-1} \subset \Gamma^1$.

We may assume that A_i is located at the origin and Γ_i lies on x_i-axes. To simplify the notation we will write $S_{i,\delta_i} = S_\delta$ and $\beta_i = \beta$, etc.

We will prove case (i) only. The proof for the other two cases is analogous.

Obviously the solution of problem (2.1) satisfies

$$\begin{aligned} &-\Delta u = f, \\ &u|_{\tilde{\Gamma}_i} = G^0|_{\tilde{\Gamma}_i}, \\ &\left.\frac{\partial u}{\partial n}\right|_{\tilde{\Gamma}_{i-1}} = G^1|_{\tilde{\Gamma}_{i-1}} \end{aligned} \tag{2.41}$$

where

$$\tilde{\Gamma}_l = \Gamma_l \cap S_\delta, \qquad l = i-1, i.$$

Let

$$\begin{aligned} &\varphi_0 \in C^\infty(R_+^1), \\ &\varphi_0(x) = 1 \quad \text{for } 0 \leqq x \leqq \tfrac{1}{2}, \\ &\varphi_0(x) = 0 \quad \text{for } x \geqq 1, \\ &\varphi_\delta(r) = \varphi_0\left(\frac{r}{\delta}\right) = \varphi(r). \end{aligned}$$

Denote

$$v = \varphi u.$$

Then, by zero extension outside S_δ, function v is defined on the infinite sector $Q = \{(r, \theta) \mid 0 < r < \infty,\ 0 < \theta < \omega\}$ and v satisfies

$$\begin{aligned} &-\Delta v = \varphi f + 2\nabla\varphi\nabla u + u\Delta\varphi = \tilde{f}, \\ &v|_{\theta=0} = \varphi G^0|_{\theta=0} = \tilde{G}^0|_{\theta=0}, \\ &\left.\frac{\partial v}{\partial n}\right|_{\theta=\omega} = \frac{1}{r}\left.\frac{\partial v}{\partial \theta}\right|_{\theta=\omega} = \varphi G^1|_{\theta=\omega} = \tilde{G}^1|_{\theta=\omega}. \end{aligned} \tag{2.42}$$

Obviously $v \in H^1(Q)$ and $\tilde{f}$, $\tilde{G}^0$, $\tilde{G}^1 = 0$ for $r > 1$. Denote by w the solution of (2.24) mentioned in Lemma 2.6. Then using Lemma 2.8 we see that $v = w$ and hence by (2.25)

$$\begin{aligned} \|v\|_{\mathscr{H}_\beta^{2,2}(S_\delta)} &\leqq C[\|\tilde{f}\|_{\mathscr{L}_\beta(Q)} + \|\tilde{G}^0\|_{\mathscr{H}_\beta^{2,2}(Q)} + \|\tilde{G}^1\|_{\mathscr{H}_\beta^{1,1}(Q)}] \\ &\leqq C\left[\|\tilde{f}\|_{\mathscr{L}_\beta(S_\delta)} + \sum_{l=0}^{1} \|\tilde{G}^l\|_{\mathscr{H}_\beta^{2-l,2-l}(S_\delta)}\right]. \end{aligned} \tag{2.43}$$

In (2.43) we have used the fact that $\varphi = 0$ for $r > \delta$. Because $\nabla\varphi = \Delta\varphi = 0$ for $0 < r < \delta/2$ and $r > \delta$ we have

$$\|\nabla\varphi\nabla u\|_{\mathscr{L}_\beta(S_\delta)} \leqq C\|u\|_{\mathscr{H}^{1,1}_\beta(S_\delta - S_{\delta/2})} \leqq C_1,$$

$$\|u\Delta\varphi\|_{\mathscr{L}_\beta(S_\delta)} \leqq C\|u\|_{\mathscr{L}_\beta(S_\delta - S_{\delta/2})} \leqq C_1.$$

Because $f \in B^0_\beta(\Omega)$, $G^i \in B^{2-i}_\beta(\Omega)$ we get immediately from (2.43)

$$\begin{aligned}\|u\|_{\mathscr{H}^{2,2}_\beta(S_{\delta/2})} &= \|v\|_{\mathscr{H}^{2,2}_\beta(S_{\delta/2})} \leqq \|v\|_{\mathscr{H}^{2,2}_\beta(S_\delta)} \\ &\leqq C_2\left\{\|f\|_{\mathscr{L}_\beta(S_\delta)} + \sum_{l=0}^{1}\|G^l\|_{\mathscr{H}^{2-l,2-l}_\beta(S_\delta)} + \|u\|_{\mathscr{H}^{1,1}_\beta(S_\delta - S_{\delta/2})}\right\}\end{aligned} \tag{2.44}$$

with C_2 dependent on δ and φ. Hence (2.40) holds for $|\alpha| = 2$. Let

$$v_k = r^k u_{r^k}, \qquad k \geqq 2. \tag{2.45}$$

Then

$$\begin{aligned}&-\Delta v_k = r^{k-2}(r^2 f)_{r^k} \quad \text{in } S_\delta, \\ &v_k|_{\theta=0} = r^k G^0_{r^k}|_{\theta=0}, \\ &\left.\frac{\partial v_k}{\partial n}\right|_{\theta=\omega} = \frac{1}{r}\left.\frac{\partial v_k}{\partial\theta}\right|_{\theta=\omega} = (r^k G^1_{r^k} + kr^{k-1}G^1_{r^{k-1}})|_{\theta=\omega}.\end{aligned} \tag{2.46}$$

Let $w_k = \varphi v_k$. Then

$$-\Delta w_k = -\varphi\Delta v_k - 2\nabla\varphi\nabla v_k - v_k\Delta\varphi$$

and

$$\begin{aligned}&w_k|_{\theta=0} = r^k G^0_{r^k}|_{\theta=0}, \\ &\left.\frac{\partial w_k}{\partial n}\right|_{\theta=\omega} = \frac{1}{r}\left.\frac{\partial w_k}{\partial\theta}\right|_{\theta=\omega} = \varphi(r^k G^1_{r^k} + kr^{k-1}G^1_{r^{k-1}})|_{\theta=\omega}.\end{aligned}$$

Hence analogously as before

$$\begin{aligned}\|v_k\|_{\mathscr{H}^{2,2}_\beta(S_{\delta/2})} \leqq C[&\|r^{k-2}(r^2 f)_{r^k}\|_{\mathscr{L}_\beta(S_\delta)} + \|v_k\|_{\mathscr{H}^{1,1}_\beta(S_\delta - S_{\delta/2})} + \|v_k\|_{\mathscr{L}_\beta(S_\delta - S_{\delta/2})} \\ &+ \|r^k G^0_{r^k}\|_{\mathscr{H}^{2,2}_\beta(S_\delta)} + \|r^k G^1_{r^k}\|_{\mathscr{H}^{1,1}_\beta(S_\delta)} + k\|r^{k-1}G^1_{r^{k-1}}\|_{\mathscr{H}^{1,1}_\beta(S_\delta)}].\end{aligned} \tag{2.47}$$

Because $f \in B^0_\beta(\Omega)$, $G^l \in B^{2-i}_\beta(\Omega)$ we have

$$\|r^{k-2}(r^2 f)_{r^k}\|_{\mathscr{L}_\beta(S_\delta)} \leqq C_3 d_3^k k!, \tag{2.48a}$$

$$\|r^k G^i_{r^k}\|_{\mathscr{H}^{2-i,2-i}_\beta(S_\delta)} \leqq C_4 d_4^k k!, \tag{2.48b}$$

$$k\|r^{k-1}G^1_{r^{k-1}}\|_{\mathscr{H}^{1,1}_\beta(S_\delta)} \leqq C_5 d_5^k k!. \tag{2.48c}$$

Using (2.39) we get

$$\|v_k\|_{\mathscr{H}^{1,1}_\beta(S_\delta - S_{\delta_2})} \leqq C d_0^{k-1}(k-1)!,$$

$$\|v_k\|_{\mathscr{H}^0_\beta(S_\delta - S_{\delta/2})} \leqq C d_0^{k-2}(k-2)!$$

with C independent of k (depending on δ). Hence

$$\|v_k\|_{\mathscr{H}^{2,2}_\beta(S_{\delta/2})} \leqq C_6 d_6^k k! \tag{2.49}$$

with C_6 and d_6 independent of k, L, D and P. Assume now by induction that (2.40) holds for $k' < k$. Then we get using (2.49)

$$\begin{aligned}\|r^k u_{r^{k+2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} &\leqq C_6 d_6^k k! + 2kLD^{k-1}(k-1)! + k(k-1)LD^{k-2}(k-2)! \\ &\leqq LD^k k!\end{aligned} \tag{2.50a}$$

provided that D and L are large enough. Further $P > 1$, e.g., $P = 2$

$$\|r^{k-1} u_{r^{k+1}\theta}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq C_6 d_6^k k + k(k-1)! LD^{k-1} P \leqq LD^k Pk! \tag{2.50b}$$

$$\|r^{k-2} u_{r^k\theta^2}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq C_6 d_6^6 k! < LK^k P^2 k!. \tag{2.50c}$$

Inequalities (2.50) yield (2.40) with $\alpha_1 \geqq 2$ and $\alpha_2 \leqq 2$.

Let us now prove (2.40) by induction with respect to α_2. Let $v = r^{\alpha_1} u_{r^{\alpha_1}\theta^{\alpha_2-2}}$, $\alpha_2 \geqq 2$, $\alpha_1 + \alpha_2 = k$, $\alpha_1 \geqq 2$. Then

$$-\Delta v = r^{\alpha_1-2}(r^2 f_{\theta^{\alpha_2-2}})_{r^{\alpha_1}}$$

and also

$$-\Delta v = -r^{\alpha_1-2} u_{r^{\alpha_1}\theta^{\alpha_2}} - (2\alpha_1+1) r^{\alpha_1-1} u_{r^{\alpha_1+1}\theta^{\alpha_2-2}} - \alpha_1^2 r^{\alpha_1-2} u_{r^{\alpha_1}\theta^{\alpha_2-2}} - r^{\alpha_1} u_{r^{\alpha_1+2}\theta^{\alpha_2-2}}.$$

Hence

$$\begin{aligned}\|r^{\alpha_1-2} u_{r^{\alpha_1}\theta^{\alpha_2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq{}& \|r^{\alpha_1-2}(r^2 f_{\theta^{\alpha_2-2}})_{r^{\alpha_1}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \\ &+ (2\alpha_1+1)\|r^{\alpha_1-1} u_{r^{\alpha_1+1}\theta^{\alpha_2-2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \\ &+ \alpha_1^2 \|r^{\alpha_1-2} u_{r^{\alpha_1}\theta^{\alpha_2-2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \\ &+ \|r^{\alpha_1} u_{r^{\alpha_1+2}\theta^{\alpha_2-2}}\|_{\mathcal{L}_\beta(S_{\delta/2})}.\end{aligned} \tag{2.51}$$

Because $f \in B_\beta^0(\Omega)$ we have

$$\|r^{\alpha_1-2}(r^2 f_{\theta^{\alpha_2-2}})_{r^{\alpha_1}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq C_3 d_3^{k-2}(k-2)!. \tag{2.52}$$

By the induction assumption

$$\|r^{\alpha_1-1} u_{r^{\alpha_1+1}\theta^{\alpha_2-2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq LD^{k-3} P^{\alpha_2-2}(k-3)!, \tag{2.53a}$$

$$\|^{\alpha_1} u_{r^{\alpha_1+2}\theta^{\alpha_2-2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq LD^{k-2} P^{\alpha_2-2}(k-2)!, \tag{2.53b}$$

$$\|r^{\alpha_1-2} u_{r^{\alpha_1}\theta^{\alpha_2-2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} \leqq LD^{k-4} P^{\alpha_2-2}(k-4)!. \tag{2.53c}$$

Hence from (2.47)

$$\begin{aligned}\|r^{\alpha_1-2} u_{r^{\alpha_1}\theta^{\alpha_2}}\|_{\mathcal{L}_\beta(S_{\delta/2})} &\leqq (k-2)![C_3 d_3^{k-2} + LD^{k-2} P^{\alpha_2-2} + LD^{k-3} P^{\alpha_2-2}(2\alpha_1+1) \\ &\qquad + LD^{k-4} P^{\alpha_2-2} \alpha_1^2] \\ &\leqq LD^{k-2} P^{\alpha_2}(k-2)!\end{aligned} \tag{2.54}$$

provided that L, D and P are sufficiently large. Similarly we can prove (2.40) for $\alpha_2 \geqq 2$, $\alpha_1 = 0, 1$.

Theorem 2.1 is proven for the case (i). The other cases are analogous.

Combining our results for every vertex we easily complete the proof. □

3. Regularity of the solution of the elliptic equation in a polygonal domain Ω. In § 2 we analyzed the problem of the regularity of the solution of the Poisson problem on a polygonal domain. In this section we will consider the general case of the elliptic equation of second order with analytic coefficients.

3.1. The problem and its basic properties. Let

$$L(u) = -\sum_{i,j=1}^{2} (a_{i,j}u_{x_i})_{x_j} + \sum_{i=1}^{2} b_i u_{x_i} + cu. \tag{3.1}$$

Let us consider the problem

$$\begin{aligned} &L(u) = f \quad \text{in } \Omega, \\ &u|_{\Gamma^0} = g^0 = G^0|_{\Gamma^0}, \\ &\left.\frac{\partial u}{\partial n_c}\right|_{\Gamma^1} = g^1 = G^1|_{\Gamma^1} \quad \text{on } \Gamma^1 \end{aligned} \tag{3.2}$$

where

$$\Gamma^0 = \bigcup_{i=\mathscr{D}} \bar{\Gamma}_i, \qquad \Gamma^1 = \Gamma - \Gamma^0,$$

and n_c is the conormal.

Let Ω be the polygonal domain in $\mathbf{R}^2$ and Γ_i be the open edge of $\partial\Omega$ (see § 2.1).

About f and g^i, $i=0, 1$ we will make the same assumptions as in Theorem 2.1 but replacing ω_i by $\omega_i^* \in (0, 2\pi]$, which will be defined later. About the operator L we will assume that

(i) $a_{i,j} = a_{j,i}$, b_i, c are analytic function on $\bar{\Omega}$,

(ii)

$$\sum_{l,j=1}^{2} a_{i,j}\xi_i\xi_j \geqq \mu_0(\xi_1^2 + \xi_2^2), \qquad \mu_0 > 0, \tag{3.3}$$

i.e., the operator is strongly elliptic.

(iii) Denote (see § 1.2)

$$H_0^1(\Omega) = \{u \in H^1(\Omega) \,|\, u = 0 \text{ on } \Gamma^0\}$$

and

$$B(u, v), \quad u \in H_0^1(\Omega), \quad v \in H_0^1(\Omega)$$

the bilinear form

$$B(u, v) = \int_\Omega (a_{i,j}u_{x_i}v_{x_j} + b_i u_{x_i} v + cuv)\, dx. \tag{3.4}$$

Assume that

$$\inf_{\substack{\|u\|_{H^1(\Omega)}=1 \\ u \in H_0^1(\Omega)}} \sup_{\substack{\|v\|_{H^1(\Omega)}=1 \\ v \in H_0^1(\Omega)}} |B(u, v)| \geqq \gamma > 0, \tag{3.5a}$$

and for any $v \in H_0^1(\Omega)$, $v \neq 0$

$$\sup_{\substack{\|u\|_{H^1(\Omega)}=1 \\ u \in H_0^1(\Omega)}} |B(u, v)| > 0. \tag{3.5b}$$

Conditions (i), (ii) guarantee the existence and uniqueness of $u \in H_0^1(\Omega)$ such that

$$B(u, v) = F(v) \tag{3.6}$$

holds for any $v \in H_0^1(\Omega)$ for any $F(v) \in H_0^1(\Omega))'$, i.e., $F(v)$ being a linear functional on $H_0(\Omega)$. In addition we have

$$\|u\|_{H^1(\Omega)} \leqq C\|F\|_{(H_0^1(\Omega))'} \tag{3.7}$$

with C independent of F.

For the proof see e.g. [4, p. 112]. Hence we have

LEMMA 3.1. *Let* $f \in \mathscr{L}_\beta(\Omega) = H_\beta^{0,0}(\Omega)$, $G^i \in H_\beta^{2-i,2-i}(\Omega)$ *and* $|\Gamma^0| \neq 0$, $\beta = (\beta_1, \cdots, \beta_M)$, $0 < \beta_i < 1$. *Then* (3.2) *has the unique solution* $u \in H^1(\Omega)$ (*in the weak sense*) *and*

$$\|u\|_{H^1(\Omega)} \leqq C[\|f\|_{\mathscr{L}_\beta(\Omega)} + \sum_{i=0,1} \|G^i\|_{H_\beta^{2-i,2-i}(\Omega)}].$$

The proof is completely analogous to the proof of Lemma 2.11, only replacing the Lax-Milgram lemma by its generalized form based on (3.5), (3.6), (3.7).

Remark 1. Condition (3.5a) and (3.5b) exclude the case when $\Gamma^0 = \varnothing$. Nevertheless this case which occurs in the case of Neumann problem and $b_i = c = 0$ can be treated in the usual way by restricting of $H_0^1(\Omega)$ to a modulo space.

LEMMA 3.2. *Let* L^0 *be the operator* (3.1) *with* $a_{i,j}^0 = a_{j,i}^0$ *constants and* $b_i = c = 0$. *Let* M *be the linear transformation*

$$\begin{aligned} &\xi_1 = (a_{1,2}^0 x_1 - a_{1,1}^0 x_2)/\sqrt{a_{1,1}^0}A, \qquad A = (a_{1,1}^0 a_{2,2}^0 - |a_{1,2}^0|^2)^{1/2}, \\ &\xi_2 = x_1/\sqrt{a_{1,1}^0} \end{aligned} \tag{3.8}$$

and $\tilde{u}(\xi_1, \xi_2) = u(x_1(\xi_1, \xi_2), x_2(\xi_1, \xi_2))$. *Then*

$$L^0 u = (-\Delta \tilde{u})$$

and the conormal n_c *in* (x_1, x_2) *transforms into the normal* n *in* (ξ_1, ξ_2). □

The lemma follows easily by simple computation.

The transformation M maps the polygonal domain Ω into the polygonal domain Ω^* with interior angles ω_i^*, $\omega_i^* = M(\omega_i)$.

Now let L be the general operator (3.1). By assumption the coefficients $a_{i,j}$ are analytic on $\bar{\Omega}$. Hence we can define mappings M_k associated to the vertices A_k with $a_{i,j}^0 = a_{i,j}(A_k)$ and set $\omega_k^* = M(\omega_k)$.

3.2. The regularity of the solution. The main theorem of this chapter is as follows.

THEOREM 3.1. *Let* $f \in B_\beta^0(\Omega)$, $g^l \in B_\beta^{3/2-l}(\Gamma^l)$, $l = 0, 1$, $\beta = (\beta_1, \cdots, \beta_M)$, $0 < \beta_i < 1$, $\beta_i > 1 - \pi/\omega_1^*$ (*respectively* $\beta_i > 1 - \pi/2\omega_i^*$ *if Dirichlet and Neumann boundary conditions are imposed on the edges* Γ_i *and* Γ_{i-1}, $\bar{\Gamma}_i \cap \bar{\Gamma}_{i-1} = A_i$) *and* $\Gamma^0 \neq \varnothing$. *Then problem* (3.1) *has a unique solution in* $H^1(\Omega)$ *and* $u \in B_\beta^2(\Omega)$.

Proof. The main idea of the proof is the same as in Theorem 2.1, namely that in the neighborhood of every vertex A_i the inequality (2.40) holds. By Theorem 5.7.1, 5.7.1′ and 6.6.1 of [16] u is analytic in Ω and on (open) Γ_i, $1 \leqq i \leqq M$.

Let the mapping M_l map Ω into Ω^* with the vertex A_l mapped into the origin and the edge Γ_l being mapped into Γ_l^* lying on the ξ_1 axis. Defining $\tilde{u}(\xi_1, \xi_2) = u(M_l^{-1}(\xi_1, \xi_2))$ we have

$$L^*(\tilde{u}) = \tilde{f} \tag{3.9}$$

where

$$L^*(\tilde{u}) = -\Delta \tilde{u} - \sum_{i,j=1}^{2} \tilde{a}_{i,j} \tilde{u}_{\xi_i \xi_j} + \sum_{j=1}^{2} \tilde{b}_j \tilde{u}_{\xi_j} + \tilde{c}\tilde{u} \tag{3.10}$$

with $\tilde{a}_{i,j}(0, 0) = 0$, and $\tilde{a}_{i,j}$, $\tilde{b}_j$ and $\tilde{c}$ are analytic functions in $\bar{\Omega}^*$.

Let $S=\{r, \theta \mid 0<r<\delta_l, 0<\theta<\omega_l^*\}$, $\delta_l=\delta<1$ and $S\subset\Omega^*$. Let us analyze in detail the case Γ_{l-1}^*, $\Gamma_l^*\subset\Gamma^{0*}$. (Let us write further Γ_l instead of Γ_l^*.) We have

$$
\begin{aligned}
&L^*(\tilde{u})=\tilde{f} \quad \text{on } S,\\
&\tilde{u}|_{\tilde{\Gamma}_l\cup\tilde{\Gamma}_{l-1}}=G^0|_{\tilde{\Gamma}_l\cup\tilde{\Gamma}_{l-1}}
\end{aligned}
\tag{3.11}
$$

where $\tilde{\Gamma}_l=\Gamma_l\cap S$. Without a loss of generality we assume that $G^0=0$ (if not we set $v=u-G^0$). We rewrite (3.10) by replacing $\tilde{u}$, $\tilde{a}_{i,j}$, $\tilde{b}_j$, $\tilde{c}$, $\tilde{f}$ by u, $a_{i,j}$, b_j, c, f, and (ξ_1, ξ_2) by (x_1, x_2). Then

$$
-\Delta u=f_1=f+\sum_{i,j=1}^{2} a_{i,j}u_{x_ix_j}-\sum_{j=1}^{2} b_ju_{x_j}-cu, \qquad u|_{\tilde{\Gamma}_l\cup\tilde{\Gamma}_{l-1}}=0. \tag{3.12}
$$

By Theorem 2.1 (see (2.44)) we have for $\beta_l>1-\pi/\omega_l^*$ and $\delta_1<\delta/2$

$$
\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}\leqq C_0(\|f_1\|_{\mathscr{L}_\beta(S_\delta)}+\|u\|_{H^1(S_\delta-S_{\delta/2})}) \tag{3.13}
$$

where for simplicity we set $\beta=\beta_l$. Since $a_{i,j}(0,0)=0$ and $a_{i,j}$ are analytic in $\bar{\Omega}^*$ we have $|a_{i,j}|\leqq C_1 r$ in $\bar{S}_\delta$ and hence

$$
\|a_{i,j}u_{x_ix_j}\|_{\mathscr{L}_\beta(S_{\sigma_1})}\leqq C_1\delta_1\|u_{x_ix_j}\|_{\mathscr{L}_\beta(S_{\delta_1})}. \tag{3.14}
$$

One has

$$
u_{x_1}=u_r\cos\theta-u_\theta\frac{\sin\theta}{r},
$$

$$
u_{x_1^2}=u_{r^2}\cos^2\theta-u_{r\theta}\frac{\sin 2\theta}{r}+\frac{1}{r^2}u_{\theta^2}\sin^2\theta+\frac{1}{r}u_r\sin^2\theta+\frac{1}{r^2}u_\theta\sin 2\theta
$$

and similar expressions for $u_{x_1x_2}$ and $u_{x_2^2}$. Using Lemma A.2 scaled to the sector S_{δ_1} we get for $|\alpha|=1$

$$
\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})}\leqq\tilde{C}_2\Big(\sum_{|\alpha'|=2}\|r^{\alpha_1'-2}\mathscr{D}^{\alpha'}u\|_{\mathscr{L}_\beta(S_{\delta_1})}+\|u\|_{H^1(S_{\delta_1})}\Big)
$$

with $\tilde{C}_2\geqq 1$ independent of δ_1.

Hence

$$
\begin{aligned}
\|u_{x_1^2}\|_{\mathscr{L}_\beta(S_{\delta_1})}&\leqq C_2\Big(\sum_{|\alpha|=2}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})}+\|u\|_{H^1(S_{\delta_1})}\Big)\\
&\leqq C_2\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}.
\end{aligned}
\tag{3.15}
$$

Analogously it can be readily proven that (3.15) holds for $u_{x_ix_j}$, $i=1, 2$.

Using (3.14) in (3.15) we get

$$
\begin{aligned}
\sum_{i,j=1}^{2}\|a_{i,j}u_{x_ix_j}\|_{\mathscr{L}_\beta(S_{\delta_1})}&\leqq C_2C_1\delta_1\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}\\
&=C_3\delta_1\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}
\end{aligned}
\tag{3.16}
$$

where C_3 is independent of u and δ_1. Let us select δ_1 so that $C_0C_3\delta_1<\frac{1}{2}$. Then we get from (3.13)-(3.16)

$$
\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}\leqq C_0[\|f\|_{\mathscr{L}_\beta(\delta)}+\|u\|_{H^1(S_\delta-S_{\delta/2})}+\|u\|_{H^2(S_\delta-S_{\delta/1})}]+C_0C_3\delta_1\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}. \tag{3.17}
$$

Hence

$$
\|u\|_{\mathscr{H}_\beta^{2,2}(S_{\delta_1})}\leqq C_4[\|f\|_{\mathscr{L}_\beta(\delta)}+\|u\|_{H^1(S_\delta)}+\|u\|_{H^2(S_\delta-S_{\delta_1})}]. \tag{3.18}
$$

Because u is analytic we have for any $|\alpha| = k$

$$\|D^\alpha u\|_{L_2(S_\delta - S_{\delta_1})} \leqq C_5 d_5^k k! \tag{3.19}$$

and we have also $u \in H^1(S_\delta)$. Hence $u \in \mathscr{H}^{2,2}_\beta(S_{\delta_1})$.

Let now $v_k = r^k u_{r^k}$, $k \geqq 1$. Then

$$\begin{aligned} -\Delta v_k &= f_k = f_k^{(1)} + f_k^{(2)} \quad \text{on } S_{\delta_1}, \\ v_k|_{\tilde{\Gamma}_l \cup \tilde{\Gamma}_{l-1}} &= 0 \end{aligned} \tag{3.20}$$

where

$$f_k^{(1)} = r^{k-2}(r^2 f)_{r^k}, \tag{3.21a}$$

$$f_k^{(2)} = r^{k-2}\left(r^2 \sum_{l,j=1}^{2} a_{i,j} u_{x_i x_j} - r^2 \sum_{j=1}^{2} b_j u_{x_j} - r^2 c u \right)_{r^k}. \tag{3.21b}$$

As before we have

$$\|v_k\|_{\mathscr{H}^{2,2}_\beta(S_{\delta_1})} \leqq C_0(\|f_k\|_{\mathscr{L}_\beta(S_\delta)} + \|v_k\|_{H^1(S_\delta - S_{\delta/2})}). \tag{3.22}$$

Since by assumption $f \in B^0_\beta(\Omega)$ there exist constants C_f, d_f such that

$$\|f_k^{(1)}\|_{\mathscr{L}_\beta(S_\delta)} \leqq C_f d_f^k d!. \tag{3.23}$$

Because the coefficients $a_{i,j}$, b_j, c are analytic in $\bar{\Omega}$ there exist constants C_6, d_6 such that for $|\alpha| = k > 0$

$$\begin{aligned} |D^\alpha a_{i,j}| &\leqq C_6 d_6^k k!, \\ |D^\alpha b_j| &\leqq C_6 d_6^k k!, \\ |D^\alpha c| &\leqq C_6 d_6^k k!. \end{aligned} \tag{3.24}$$

Hence for $0 \leqq k - m \leqq 2$

$$|(r^2 a_{i,j})_{r^{k-m}}| \leqq \tilde{C}_6 r^{(2-k+m)} d_6^{(2-k+m)}$$

and for $k - m \geqq 2$

$$|(r^2 a_{i,j})_{r^{k-m}}| \leqq \tilde{C}_6 d_6^{k-m}(k-m)!.$$

Therefore for $i, j = 1, 2$

$$\begin{aligned} \|r^{k-2}(r^2 a_{i,j} u_{x_i x_j})_{r^k}\|_{\mathscr{L}_\beta(S_{\delta_1})} &\leqq \sum_{m=0}^{k} \binom{k}{m} \|r^{k-2}(r^2 a_{i,j})_{r^{k-m}} (u_{x_i x_j})_{r^m}\|_{\mathscr{L}_\beta(S_{\delta_1})} \\ &\leqq C_1 \delta_1 \|r^k u_{x_i x_j r^k}\|_{\mathscr{L}_\beta(S_{\delta_1})} \\ &\quad + \sum_{m=0}^{k-1} \binom{k}{m} \tilde{C}_6 d_6^{k-m}(k-m)! \|r^{k-2+\xi(m,k)} u_{x_i x_j r^m}\|_{\mathscr{L}_\beta(S_{\delta_1})} \end{aligned} \tag{3.25}$$

where

$$\begin{aligned} \xi(m, k) &= 2 - k + m \quad \text{for } 0 < k - m \leqq 2, \\ \xi(m, k) &= 0 \qquad\qquad\ \ \text{for } k - m > 2. \end{aligned}$$

Obviously

$$|\mathscr{D}^2 v_k| \leqq r^k |\mathscr{D}^2 u_{r^k}| + 2k|r^{k-1} u_{r^{k+1}}| + k(k-1)|r^{k-2} u_{r^k}| \tag{3.26}$$

and by (A.10) of Lemma A.4 we get

$$\begin{aligned}(3.27)\qquad &\|r^k u_{x_i x_j r^k}\|_{\mathscr{L}_\beta(S_{\delta_1})}\\ &\leqq \|v_k\|_{\mathscr{H}^{2,2}_\beta(S_{\delta_1})} + C_7 k!\left(\sum_{\substack{2\leqq|\alpha|\leqq k+1\\ 0\leqq\alpha_2\leqq 2}} \frac{k-|\alpha|+3}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})} + \|u\|_{H^1(S_{\delta_1})}\right).\end{aligned}$$

For $m\leqq k-1$, by (A.11) of Lemma A.4

$$\begin{aligned}(3.28)\qquad &\|r^{k-2+\xi(m,k)} u_{x_i x_j r^m}\|_{\mathscr{L}_\beta(S_{\delta_1})}\\ &\leqq C_7 m!\left(\sum_{\substack{2\leqq|\alpha|\leqq m+2\\ 0\leqq\alpha_2\leqq 2}} \frac{(m-|\alpha|+3)}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})} + \|u\|_{H^1(S_{\delta_1})}\right).\end{aligned}$$

Hence

$$\begin{aligned}(3.29)\qquad &\sum_{m=0}^{k-1}\binom{k}{m}\|r^{k-2}(r^2 a_{ij})_{r^{k-m}}(u_{x_i x_j})_{r^m}\|_{\mathscr{L}_\beta(S_{\delta_1})}\\ &\leqq C_7 k!\sum_{m=0}^{k-1} d_7^{k-m}\left(\sum_{2\leqq|\alpha|\leqq m+2}\frac{(m-|\alpha|+3)}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})} + \|u\|_{H^1(S_{\delta_1})}\right).\end{aligned}$$

Similarly

$$\begin{aligned}(3.30)\qquad &\left\|r^{k-2}\left(r^2\sum_{j=1}^{2} b_j u_{x_j} + r^2 cu\right)_{r^k}\right\|_{\mathscr{L}_\beta(S_{\delta_1})}\\ &\leqq C_7 k!\left(\sum_{m=0}^{k} d_7^k \sum_{\substack{2\leqq|\alpha|\leqq m+1\\ 0\leqq\alpha_2\leqq 1}}\frac{1}{(|\alpha|-1)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})} + \|u\|_{H^1(S_{\delta_1})}\right).\end{aligned}$$

Thus

$$\begin{aligned}(3.31)\qquad &\|v_k\|_{\mathscr{H}^{2,2}_\beta(S_{\delta_1})} \leqq C_0 C_1\delta_1\|v^k\|_{\mathscr{H}^{2,2}_\beta(S_{\delta_1})}\\ &+ C(\delta)k!\left[\left(\sum_{\substack{2\leqq|\alpha|\leqq k+1\\ 0\leqq\alpha_2\leqq 2}}\frac{k-|\alpha|+3}{(|\alpha|-2)!}\|r_1^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})}\right.\right.\\ &+ \|u\|_{H^1(S_{\delta_1})} + \sum_{m=0}^{k-1} d_7^{k-m}\sum_{\substack{2\leqq|\alpha|\leqq m+2\\ 0\leqq\alpha_2\leqq 2}}\frac{(m-|\alpha|+3)}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})}\\ &+ \sum_{m=0}^{k} d_7^{k-m}\sum_{\substack{2\leqq|\alpha|\leqq m+1\\ 0\leqq\alpha_2\leqq 1}}\frac{1}{(|\alpha|-1)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_{\delta_1})}\\ &\left.\left. + C_f d_f^k k! + \|u\|_{H^{k+2}(S_\delta - S_{\delta_1})}\right)\right].\end{aligned}$$

Assuming by induction that (2.40) holds for $\alpha_2\leqq 2$ and $|\alpha|\leqq k-1$ and realizing that for $C_0C_1\delta_1<\frac{1}{2}$ we get (2.40) for $|\alpha|=k$, $\alpha_2\leqq 2$ provided that L and D are sufficiently large.

The same argument as has been used in § 2 yields (2.40) for $\alpha_2 > 2$. So far we have assumed that $(\Gamma_{l-1} \cup \Gamma_l) \subset \Gamma^0$. In the case $\Gamma_l \subset \Gamma^1$, $\Gamma_{l-1} \subset \Gamma^0$ we proceed analogously. We have

$$-\Delta u = f + \sum_{i,j=1}^{2} a_{i,j} u_{x_i x_j} - \sum_{j=1}^{2} b_j u_{x_j} - cu = f_1 \quad \text{on } S_{\delta_1}, \quad u|_{\tilde{\Gamma}_{l-1}} = 0,$$

$$\left.\frac{\partial u}{\partial n}\right|_{\tilde{\Gamma}_l} = \left[\frac{\partial u}{\partial n_c} - (a_{1,2} u_{x_1} + a_{2,2} u_{x_2})\right]\Bigg|_{\tilde{\Gamma}_l}$$

$$= G^1|_{\tilde{\Gamma}_l} - (a_{1,2} u_{x_1} + a_{2,2} u_{x_2})| = \tilde{G}^1|_{\tilde{\Gamma}_l}$$

with $a_{i,j}(0,0) = 0$. By (2.44) we have

$$\|u\|_{\mathscr{K}^{2,2}_{\beta}(S_{\delta_1})} \leqq C(\|f_1\|_{\mathscr{L}_{\beta}(S_\delta)} + \|G^1\|_{\mathscr{K}^{1,1}_{\beta}(S_\delta)} + \delta_1 \|u\|_{\mathscr{K}^{2,2}_{\beta}(S_{\delta_1})} + \|u\|_{H^1(S_\delta)} + \|u\|_{H^2(S_\delta - S_{\delta_1})})$$

and the proof is very similar as before. The same arguments hold for the case $(\Gamma_l \cup \Gamma_{l-1}) \subset \Gamma^1$. Combining the results for every vertex we get our theorem. □

Remark 2. In the proof that $u \in B^2_{\beta}(\Omega)$ we have only assumed that the solution exists. The other conditions, namely (3.4) and (3.5), only guarantee this existence.

We have assumed that the coefficients $a_{i,j}$, b_j, c are analytic on $\bar{\Omega}$. This assumption can be weakened. For example, we can assume that $a_{i,j}$, b_j, c are analytic on $\bar{\Omega} - \cup_{j=1}^{M} A_j$ and in the neighborhood of A_l, $l = 1, \cdots, M$

$$|D^\alpha (a_{i,j} - a_{i,j}(A_l))| \leqq C d^k k! r_l^{\varepsilon_l^a - k},$$

$$|D^\alpha b_j| \leqq C d^k k! r_l^{\varepsilon_l^b - k - 1},$$

$$|D^\alpha c| \leqq C d^k k! r_l^{\varepsilon_l^c - k - 2},$$

with arbitrary $\varepsilon_l^a > 0$, $\varepsilon_l^b > 0$, $\varepsilon_l^c > 0$, $\varepsilon_l^c + \beta_l > 1$, and $k = |\alpha|$. Nevertheless we will not go in further detail although this case plays an important role when nonlinear problem is studied.

4. Appendix.

LEMMA A.1. *One has the inequality*

$$\text{(A.1)} \qquad \int_0^1 t^{\alpha-2}[z(t) - a]^2 \, dt \leqq C(\alpha) \int_0^1 t^\alpha \left(\frac{dz}{dt}\right)^2 dt, \qquad \alpha \neq 1$$

where for $\alpha < 1$, $z(t)$ *is continuous on* $(0, 1]$, *and* $a = z(0)$; *for* $\alpha > 1$, $z(t)$ *is continuous on* $(0, 1]$, *and* $a = z(1)$.

Proof. For $\alpha < 1$, we have by Theorem 2.53 of [13]

$$\int_0^1 s^{-2} |w(s)|^2 \, ds \leqq C\left[\int_0^1 |w'(s)|^2 \, ds + w^2(1)\right], \qquad w(0) = 0.$$

Because $|w(1)|^2 \leqq \int_0^1 |w'(s)|^2 \, ds$ we have

$$\int_0^1 s^{-2} |w(s)|^2 \, ds \leqq C \int_0^1 |w'(s)|^2 \, ds.$$

Setting $w(t) = z(t) - z(0)$ and $s = t^{1-\alpha}$ we get

$$\int_0^1 t^{\alpha-2} (z(t) - z(0))^2 \, dt \leqq C \int_0^1 t^\alpha |z'(t)|^2 \, dt.$$

For $\alpha > 1$ we use Theorem 2.53 of [13]

$$\int_0^\infty s^{-2} |w(s)|^2 \, ds \leqq 4 \int_0^\infty |w'(s)|^2 \, ds, \qquad w(0) = 0.$$

Setting $t = s^{1/(1-\alpha)}$, $w(s) = z(s^{1/(1-\alpha)}) - z(1)$ for $s > 1$ and $w(s) = 0$ for $s \leqq 1$. Then we get (A.1).

LEMMA A.2. *Let* $S_\delta = \{r, \theta \mid 0 < r < \delta, 0 < \theta < \omega\}$. *Then for* $0 < \beta < 1$:

(i)

$$\|r^{-1}u\|^2_{\mathscr{L}_\beta(S_1)} \leqq C\left[\sum_{|\alpha|=1} \|r^{\alpha_1 - 1}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S_1)} + \|u\|^2_{L_2(S_1 - S_{1/2})}\right]; \tag{A.2}$$

(ii) *for* $|\alpha| = 1$

$$\|r^{\alpha_1 - 2}\mathscr{D}^\alpha u\|^2_{\mathscr{L}_\beta(S_1)} \leqq C\left[\sum_{|\alpha'|=2} \|r^{\alpha_1' - 2}\mathscr{D}^{\alpha'} u\|^2_{\mathscr{L}_\beta(S_1)} + \|u\|^2_{H^1(S_1)}\right]. \tag{A.3}$$

Proof. The proof is similar to the one of [5].

(1) Let

$$\bar{u}(r) = \frac{1}{\omega}\int_0^\omega u(r, \theta)\, d\theta;$$

then

$$\bar{u}_r(r) = \frac{1}{\omega}\int_0^\omega u_r(r, \theta)\, d\theta$$

and hence

$$\int_0^1 r^{2\beta+1}|\bar{u}_r(r)|^2\, dr \leqq C\|u_r\|^2_{\mathscr{L}_\beta(S_1)}.$$

Using (A.1) we get

$$\int_0^1 r^{2\beta-1}|\bar{u}(r) - a|^2\, dr \leqq C\|u_r\|^2_{\mathscr{L}_\beta(S_1)}$$

where $a = \bar{u}(1)$, and by the imbedding theorem

$$|a|^2 \leqq C\int_{1/2}^1 (\bar{u}^2 + \bar{u}_r^2)\, dr.$$

Hence

$$\int_0^1 r^{2\beta-1}|\bar{u}(r)|^2\, dr \leqq C[\|u_r\|^2_{\mathscr{L}_\beta(S_1)} + \|u\|^2_{L_2(S_1 - S_{1/2})}]$$

and

$$\|r^{-1}\bar{u}\|^2_{\mathscr{L}_\beta(S_1)} \leqq C[\|u_r\|^2_{\mathscr{L}_\beta(S_1)} + \|u\|^2_{L_2(S_1 - S_{1/2})}]. \tag{A.4}$$

Further for almost all r, φ we get

$$u(r, \varphi) - u(r, \psi) = \int_\psi^\varphi u_\theta(r, \theta)\, d\theta$$

and therefore

$$\begin{aligned} |u(r, \ ) - \bar{u}(r)| &= |\omega^{-1}\int_0^\omega d\psi \int_\psi^\varphi u_\theta(r, \theta)\, d\theta| \\ &\leqq C\left[\int_0^\omega |u_\theta(r, \theta)|^2\, d\theta\right]^{1/2} \end{aligned}$$

and

$$\int_0^\omega |u(r,\varphi)-\bar{u}(r)|^2\,d\theta \leqq C\int_0^\omega |u_\theta(r,\theta)|^2\,d\theta.$$

Hence

$$\int_{S_1} r^{2\beta-2}|u(r,\varphi)-\bar{u}(r)|^2 r\,dr\,d\varphi \leqq C\int_{S_1} r^{2\beta-2}|u_\theta(r,\theta)|^2 r\,dr\,d\theta. \tag{A.5}$$

Combining (A.4) and (A.5) we get (A.2).

(2) Let $v=u_r$. Then using (A.2) we have

$$\|r^{-1}u_r\|^2_{\mathscr{L}_\beta(S_1)} \leqq C[\|u_{rr}\|^2_{\mathscr{L}_\beta(S_1)}+\|r^{-1}u_{r\theta}\|^2_{\mathscr{L}_\beta(S_1)}+\|u_r\|^2_{L_2(S_1-S_{1/2})}]$$

which is (A.3) for $\alpha_1=1$, $\alpha_2=0$. Now let $v=u_\theta$ and repeat our argument. Let

$$\bar{v}^2(r)=\frac{1}{\omega}\int_0^\omega v(r,\theta)\,d\theta.$$

Then

$$\bar{v}^2(r) \leqq C\int_0^\omega v^2(r,\theta)\,d\theta,$$

$$\bar{v}'(r)=\frac{d\bar{v}}{dr}(r)=\frac{1}{\omega}\int_0^\omega u_{r\theta}\,d\theta$$

and

$$\bar{v}'^2(r) \leqq C\int_0^\omega (u^2_{r\theta})\,d\theta.$$

Hence

$$\int_0^1 \bar{v}'^2(r)r^{-1+2\beta}\,dr \leqq C\int_0^1\int_0^\omega (u^2_{r\theta})r^{-2+2\beta}r\,dr\,d\theta$$
$$\leqq C\sum_{|\alpha'|=2}\|r^{\alpha_1'-2}\,^{\alpha'}u\|^2_{\mathscr{L}_\beta(S_1)}<\infty,$$

and therefore $\bar{v}(r)$ is continuous on $[0,1]$. We also have

$$\int_0^1 \bar{v}^2(r)\frac{1}{r}\,dr \leqq \|u\|^2_{H^1(S_1)}<\infty$$

and hence $\bar{v}(0)=0$.

Using now Lemma A.1 we have

$$\int_0^1 r^{2\beta-3}|\bar{v}(r)|^2\,dr \leqq C\int_0^1 r^{2\beta-1}|\bar{v}_r(r)|^2\,dr$$
$$\leqq C\|r^{-1}u_{r\theta}\|^2_{\mathscr{L}_\beta(S_1)}$$

and hence

$$\|r^{-2}\bar{v}^2(r)\|_{\mathscr{L}_\beta(S_1)} \leqq C\|r^{-1}u_{r\theta}\|_{\mathscr{L}_\beta(S_1)}. \tag{A.6}$$

Analogously as before

$$\int_0^\omega |v(r,\varphi)-\bar{v}(r)|^2\,d\varphi \leqq C\int_0^\omega |v_\theta|^2\,d\varphi$$
$$\leqq C\int_0^\omega |u_{\theta\theta}|^2\,d\varphi$$

and

$$\int_{S_1} r^{2(\beta-2)}|v(r,\varphi)-\bar{v}(r)|^2 r\,dr\,d\varphi \leqq C\int_{S_1} r^{2(\beta-2)}|u_{\theta\theta}|^2 r\,dr\,d\varphi. \tag{A.7}$$

Combining (A.6) and (A.7) we get (A.3).

LEMMA A.3. *For $0<\beta<1$ one has*

$$\|r^{-1}u\|^2_{\mathscr{L}_\beta(S_1)} \leqq C\left[\sum_{|\alpha|=1}\|D^\alpha u\|^2_{\mathscr{L}_\beta(S_1)}+\|u\|^2_{L_2(S_1-S_{1/2})}\right]. \tag{A.8}$$

For $|\alpha|=1$

$$\|r^{-1}D^\alpha u\|_{\mathscr{L}_\beta(S_1)} \leqq C\left[\sum_{|\alpha'|=2}\|D^{\alpha'}u\|^2_{\mathscr{L}_\beta(S_1)}+\|u\|^2_{H^1(S_1-S_{1/2})}\right]. \tag{A.9}$$

Proof. Because

$$\sum_{|\alpha|=1}\|D^\alpha u\|^2_{\mathscr{L}_\beta(S_1)}=\sum_{|\alpha|=1}\|r^{\alpha_1-1}\mathscr{D}^\alpha u\|^2_{\mathscr{L}_\beta(S_1)},$$

(A.8) follows immediately from (A.2). Let $v=D^\alpha u$. Then (A.9) follows immediately from (A.8). □

LEMMA A.4. *Let $S=\{r,\theta\,|\,0<r<\delta, 0<\theta<\omega\}$. Then for $k>0$, $i,j=1,2$*

$$\begin{aligned}\|r^k u_{x_ix_jr^k}\|_{\mathscr{L}_\beta(S)} &\leqq \|r^k|\mathscr{D}^2 u_{r^k}|\,\|_{\mathscr{L}_\beta(S)}\\ &\quad + Ck!\left[\sum_{\substack{2\leqq|\alpha|\leqq k+1\\ 0\leqq\alpha_2\leqq 2}}\frac{(k-|\alpha|+3)}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S)}+\|u\|_{H^1(S)}\right]\end{aligned} \tag{A.10}$$

and for $m\leqq k-1$, $k\geqq 1$, $i,j=1,2$

$$\|r^{k-2+\xi(m,k)}u_{x_ix_jr^m}\|_{\mathscr{L}_\beta(S)} \leqq Cm!\left(\sum_{\substack{2\leqq|\alpha|\leqq m+2\\ 0\leqq\alpha_2\leqq 2}}\frac{m-|\alpha|+3}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S)}+\|u\|_{H^1(S)}\right) \tag{A.11}$$

where

$$\begin{aligned}&\xi(m,k)=2-k+m \quad \text{for } 0<k-m\leqq 2,\\ &\xi(m,k)=0 \quad\qquad\ \text{for } k-m>2\end{aligned}$$

and C is independent of u, but depends on δ and ω.

Proof. We will prove (A.10) only for $i=j=1$. Proof of the other two cases is completely analogous. We have

$$u_{x_1^2}=u_{r^2}\cos^2\theta-u_{r\theta}\frac{\sin 2\theta}{r}+\frac{1}{r^2}u_{\theta^2}\sin^2\theta+\frac{1}{r}u_r\sin^2\theta+\frac{1}{r^2}u_\theta\sin 2\theta.$$

Hence

$$\begin{aligned}r^k u_{x_1^2r^k} &= r^k\left(u_{r^{2+k}}\cos^2\theta-u_{r^{k+1}\theta}\frac{\sin 2\theta}{r}+\frac{1}{r^2}u_{r^k\theta^2}\sin^2\theta\right)\\ &\quad -\sin 2\theta\sum_{l=0}^{k-1}(-1)^{k-l}\binom{k}{l}(k-l)!\,r^{l-1}u_{r^{l-1}\theta}\\ &\quad +\sin^2\theta\sum_{l=0}^{k-1}(-1)^{k-l}\binom{k}{l}(k-l+1)!\,r^{l-2}u_{r^l\theta^2}\\ &\quad +\sin^2\theta\sum_{l=0}^{k}(-1)^{k-l}\binom{k}{l}(k-l)!\,r^{l-1}u_{r^{l+1}}\end{aligned} \tag{A.12}$$

$$+\sin^2\theta \sum_{l=0}^{k} (-1)^{k/l}\binom{k}{l}(k-l+1)!\, r^{l-2}u_{r^l\theta}$$

which yields

$$\begin{aligned}
\|r^k u_{x_1^2 r^k}\|_{\mathscr{L}_\beta(S)} &\leqq \|r^k|\mathscr{D}^2 u_{r^k}|\,\|_{\mathscr{L}_\beta(S)} \\
&\quad + k!\Bigg(\sum_{l=0}^{k-1}\frac{1}{l!}\|r^{l-1}u_{r^{l+1}\theta}\|_{\mathscr{L}_\beta(S)} + \sum_{l=0}^{k-1}\frac{(k-l+1)}{l!}\|r^{l-2}u_{r^l\theta^2}\|_{\mathscr{L}_\beta(S)} \\
&\quad + \sum_{l=0}^{k-1}\frac{1}{l!}\|r^{l-1}u_{r^{l+1}}\|_{\mathscr{L}_\beta(S)} + \sum_{l=0}^{k-1}\left(\frac{k-l+1}{l!}\right)\|r^{l-2}u_{r^l\theta}\|_{\mathscr{L}_\beta(S)}\Bigg) \qquad \text{(A.13)} \\
&\leqq \|r^2|\mathscr{D}^2 u_{r^k}|\,\|_{\mathscr{L}_\beta(S)} + Ck!\Bigg(\sum_{\substack{2\leqq|\alpha|\leqq k+1 \\ \alpha_2\geqq 2}}\frac{(k-|\alpha|+3)}{(|\alpha|-2)!}\|r^{\alpha_1-2}\mathscr{D}^\alpha u\|_{\mathscr{L}_\beta(S)} \\
&\quad + \|r^{-1}u_r\|_{\mathscr{L}_\beta(S)} + \|r^{-2}u_\theta\|_{\mathscr{L}_\beta(S)}\Bigg).
\end{aligned}$$

Equation (A.13) combined with Lemma 2 yields (A.10).

The proof of (A.11) is quite analogous. □

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York-San Francisco-London, 1979.

[2] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimate near the boundary for solution of elliptic differential equations satisfying general boundary condition*, I, Comm. Pure Appl. Math., 17 (1964), pp. 35-92.

[3] M. S. AGRANOVIC AND M. I. VIŠIK, *Elliptic boundary value problem depending on a parameter*, Dokl. Akad. Nauk. SSSR, 149 (1963), pp. 223-226 (Soviet Math. Dokl., 4 (1964), pp. 325-329).

[4] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972 pp. 3-359.

[5] I. BABUŠKA AND M. R. DORR, *Error estimates for the combined h and p version of finite element method*, Numer. Math., 37 (1981), pp. 252-277.

[6] I. BABUŠKA, *The p and h-p version of the finite element method, the state of the art*, Proc. Finite Element Workshop, R. Voigt, ed., Springer, New York, 1987.

[7] I. BABUŠKA, R. B. KELLOGG AND J. PITKARANTA, *Direct and inverse error estimates for finite element method*, SIAM J. Numer. Anal., 18 (1981), pp. 515-545.

[8] L. BERGH AND J. LOFTSTROM, *Interpolation Spaces*, Springer-Verlag, Berlin-Heidelberg-New York, 1976.

[9] I. M. GELFAND AND G. E. SHILOV, *Generalized Functions*, Vol. 2, Academic Press, New York, 1964.

[10] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman Publishing, Boston, MA, 1985.

[11] B. GUO AND I. BABUŠKA, *The h-p version of finite element method. Part* 1: *The basic approximation results. Part* 2: *General results and applications*, Comp. Mech., 1 (1986), pp. 21-41, 203-220.

[12] B. GUO, *The h-p version of finite element method in two dimensions—the mathematical theory and computational experience*, Ph.D. dissertation, Univ. of Maryland, College Park, MD, 1985.

[13] G. H. HARDY, J. E. LITTLEWOOD AND G. POLYA, *Inequality*, 2nd ed., Cambridge Univ. Press, Cambridge, 1952.

[14] V. A. KONDRAT'EV, *Boundary value problem for elliptic equations in domain with conic or angular points*, Trans. Moscow Math. Soc., (1967), pp. 227-313.

[15] V. A. KONDRAT'EV AND O. A. OLEINIK, *Boundary value problems for partial differential equations in nonsmooth domains*, Russian Math. Surveys, 38 (1983), pp. 1-86.

[16] C. B. MORREY, *Multiple Integrals in Calculus of Variations*, Springer-Verlag, Berlin-Heidelberg-New York, 1966.

[17] B. A. SZABÓ, *PROBE, Theoretical Manual*, Noetic Technologies Corp., St. Louis, MO, 1985.

[18] ———, *Implementation of a finite element software system with h and p-extension capabilities*, Finite Elements in Analysis and Design, 2 (1986), pp. 177-194.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

015

A GENERALIZED DIRICHLET PRINCIPLE FOR SECOND ORDER NONSELFADJOINT ELLIPTIC OPERATORS*

ROSS G. PINSKY†

Abstract. The classical Dirichlet principle states that if the spectrum of $L=\frac{1}{2}\Delta - V$ on a smooth bounded domain $D\subset R^n$ with the Dirichlet boundary condition on ∂D is negative, then the unique solution ϕ_0 of $L\phi_0=0$ in D with $\phi_0=f$ on ∂D, for a smooth function f, minimizes a certain energy integral. This may be easily extended to the operator $L=\frac{1}{2}\nabla\cdot a\nabla + a\nabla Q\cdot\nabla - V$. Now L is selfadjoint with respect to the density e^{2Q}. In this paper, we generalize this result to nonselfadjoint operators on bounded domains. We consider the solution ϕ_0 to $L\phi_0=(\frac{1}{2}\nabla\cdot a\nabla + b\cdot\nabla - V)\phi_0=0$ in D and $\phi_0=f\geqq 0$ on ∂D under the assumption $\operatorname{Re}(\sigma(L))<0$ for L with the Dirichlet boundary condition on ∂D.

Key words. second order nonselfadjoint elliptic operator, Dirichlet principle, mini-max formula

AMS(MOS) subject classification. 35J

1. Introduction. Consider the Dirichlet problem in a smooth bounded domain D, that is,

$$\frac{1}{2}\Delta u = 0 \text{ in } D \text{ and } u=f \text{ on } \partial D \text{ for } f\in W^{1,2}(D). \tag{1.1}$$

The classical Dirichlet or energy principle states that

$$\lambda_0 = \inf_{\substack{\phi=f \text{ on } \partial D \\ \phi\in W^{1,2}(D)}} \int_D \frac{1}{2}|\nabla\phi|^2\,dx \quad \text{is attained uniquely at } \phi=\phi_0$$

where ϕ_0 is the unique solution to (1.1). More generally, if V is smooth and the lead eigenvalue of the operator $\frac{1}{2}\Delta - V$ with the Dirichlet boundary condition on ∂D is negative, then

$$\lambda_0 = \inf_{\substack{\phi=f \text{ on } \partial D \\ \phi\in W^{1,2}(D)}} \int (\tfrac{1}{2}|\nabla\phi|^2 + V\phi^2)\,dx \text{ is attained uniquely at } \phi=\phi_0$$

where ϕ_0 uniquely solves $\frac{1}{2}\Delta\phi_0 - V\phi_0=0$ in D with $\phi_0=f$ on ∂D. In fact, if Q is a smooth function and $a(x)$ is a positive definite matrix at each $x\in R^n$ with smooth entries a_{ij}, then the above result can easily be extended to the case $L=\frac{1}{2}\nabla\cdot a\nabla + a\nabla Q\cdot\nabla - V\equiv L_0 - V$. Now the Dirichlet principle states that if the lead eigenvalue of L with the Dirichlet boundary condition is negative, then

$$\lambda_0 = \inf_{\substack{\phi=f \text{ on } \partial D \\ \phi\in W^{1,2}(D)}} \left[\int_D \left(\frac{1}{2}\nabla\phi a\nabla\phi + V\phi^2\right) e^{2Q}\,dx\right] \tag{1.2}$$

is attained uniquely at $\phi=\phi_0$ where ϕ_0 uniquely solves $L\phi_0=0$ in D with $\phi_0=f$ on ∂D. Now, if we replace the first order term, $a\nabla Q\cdot\nabla$, by $b\cdot\nabla$ where $a^{-1}b$ is not a gradient, then the problem is no longer selfadjoint and although a unique solution to the Dirichlet problem still exists, it is not representable as the solution to a simple variational problem as above.

In this paper we will extend this variational principle to the nonselfadjoint case. It turns out that in the nonselfadjoint case, one must consider a mini-max variational formula rather than a standard one.

* Received by the editors July 2, 1986; accepted for publication February 9, 1987.
† Department of Mathematics, Technion, Israel Institute of Technology, Haifa 32000, Israel.

In order to see how the minimax formula was arrived at, we need to discuss briefly the classical Rayleigh-Ritz formula and its generalization to nonselfadjoint operators. The classical Rayleigh-Ritz variational formula for the largest eigenvalue of the operator $L=\frac{1}{2}\Delta - V$ for a smooth function V on a smooth domain $D\subset R^n$ with the Dirichlet (Neumann) boundary condition on ∂D may be given as follows: the largest eigenvalue λ_0 satisfies

$$\lambda_0 = \inf_{\substack{\|\phi\|=1 \\ \phi=0 \text{ on } \partial D \\ \phi\in W^{1,2}(D)}} \int_D \left(\frac{1}{2}|\nabla\phi|^2 + V\phi^2\right) dx \qquad \left(\lambda_0 = \inf_{\substack{\|\phi\|=1 \\ \phi\in W^{1,2}(D)}} \int_D \left(\frac{1}{2}|\nabla\phi|^2 + V\phi^2\right) dx\right),$$

where $\|\phi\| = (\int_D \phi^2\, dx)^{1/2}$. Furthermore, the infimum is attained uniquely at ϕ_0, where ϕ_0 is the eigenfunction corresponding to λ_0. In fact, again, if Q is a smooth function and $a(x)$ is a positive definite matrix at each $x\in R^n$ with smooth entries a_{ij}, then the above result can easily be extended to the case $L=\frac{1}{2}\nabla\cdot a\nabla + a\nabla Q\cdot\nabla - V \equiv L_0 - V$ with either the Dirichlet or the conormal boundary condition, $na\nabla\phi = 0$ on ∂D, where n is the outward unit normal. L is now selfadjoint with respect to the density e^{2Q}, and the variational formula is given by

$$\lambda_0 = \inf_{\substack{\|\phi\|=1 \\ \phi=0 \text{ on } \partial D \\ \phi\in W^{1,2}(D)}} \int_D \left(\frac{1}{2}\nabla\phi a\nabla\phi + V\phi^2\right) e^{2Q}\, dx \quad \text{in the Dirichlet case,}$$

(1.3)

$$\lambda_0 = \inf_{\substack{\|\phi\|=1 \\ \phi\in W^{1,2}(D)}} \int_D \left(\frac{1}{2}\nabla\phi a\nabla\phi + V\phi^2\right) e^{2Q}\, dx \quad \text{in the conormal case,}$$

where $\|\phi\| = (\int_D \phi^2\, e^{2Q}\, dx)^{1/2}$. The infimum is attained uniquely at ϕ_0, where ϕ_0 is the eigenfunction corresponding to λ_0. As before, if we change the first order term in L_0 from $a\nabla Q\cdot\nabla$ to $b\cdot\nabla$, where $a^{-1}b$ is not a gradient, then the operator L is no longer selfadjoint. However, L is still semibounded and has a compact resolvent; consequently the Krein-Rutman theory of positive operators [4] guarantees that $\sup(\mathrm{Re}\,(\sigma(L)))$ occurs at a real eigenvalue. Yet the classical theory does not give a variational formula for the eigenvalue.

In the nonselfadjoint setting, Donsker and Varadhan [1] and Holland [3] have obtained for the Dirichlet case and the conormal case respectively, a mini-max variational formula for $\lambda_0 = \sup(\mathrm{Re}\,(\sigma(L)))$. Now, in the selfadjoint case, comparing (1.2) and (1.3), one sees that the same functional, $\int(\frac{1}{2}\nabla\phi a\nabla\phi + V\phi^2)\, e^{2Q}\, dx$, is varied in both the Rayleigh-Ritz formula and the Dirichlet principle. The only difference is the particular subdomain of $W^{1,2}(D)$ over which the variation is taken. Thus, to generalize the Dirichlet principle, we were led to consider the functional that was obtained by Donsker and Varadhan and by Holland, and then to find the appropriate boundary conditions.

2. Statement and proof of theorem. Let $L = L_0 - V$ with $L_0 = \frac{1}{2}\nabla\cdot a\nabla + b\cdot\nabla$ in a bounded domain $D\subset R^n$ where $a(x)$ is a positive definite $n\times n$ matrix for all $x\in\bar{D}$ with entries $a_{ij}\in C^{1,\alpha}(\bar{D})$, b is an n-vector with components $b_i\in C^{1,\alpha}(\bar{D})$, $V\in C^{\alpha}(\bar{D})$ and ∂D is a $C^{2,\alpha}$-boundary. Let $\tilde{L} = \frac{1}{2}\nabla\cdot a\nabla - b\cdot\nabla - \nabla\cdot b - V$ be the formal adjoint to L. We will assume that the spectrum of the operator L with the Dirichlet boundary condition on ∂D satisfies $\mathrm{Re}\,(\sigma(L)) < 0$. Since $\sigma(\tilde{L}) = \sigma(L)$, we also have $\mathrm{Re}\,(\sigma(\tilde{L})) < 0$. Then, since $Lu = 0$ with $u = 0$ on ∂D and $\tilde{L}\tilde{u} = 0$ with $\tilde{u} = 0$ on ∂D have only the trivial solution, there exist unique solutions in $C^{2,\alpha}(\bar{D})$ of $Lu = 0$ in D with $u = f$ on ∂D and of $\tilde{L}\tilde{u} = 0$ in D with $\tilde{u} = \tilde{f}$ on ∂D, for each f and each $\tilde{f}$ in $C^{2,\alpha}(\bar{D})$ [2, Thm. 6.15]. In

particular, let ϕ_0 solve $L\phi_0 = 0$ in D with $\phi_0 = f$ on ∂D and let $\tilde{\phi}_0$ solve $\tilde{L}\tilde{\phi}_0 = 0$ in D with $\tilde{\phi}_0 = fe^{2k}$ on ∂D, where $k \in C^{2,\alpha}(\bar{D})$. From here on, we will assume that $f \geqq 0$ and $f \not\equiv 0$. Define $\partial D_1 = \partial D \cap \{f = 0\}$. We have

PROPOSITION 2.1. (a) *$\phi_0 > 0$ in D and $\tilde{\phi}_0 > 0$ in D;*

(b) *$\nabla\phi_0 \cdot n < 0$ on ∂D_1 and $\nabla\tilde{\phi}_0 \cdot n < 0$ on ∂D_1, where n is the outward unit normal.*

Proof. This generalized maximum principle follows from Theorem 10 in [7], Theorem 6.15 in [2] and the fact that $\operatorname{Re}(\sigma(L)) = \operatorname{Re}(\sigma(\tilde{L})) < 0$.

Proposition 2.1 gives us the following.

PROPOSITION 2.2. $(\phi_0\tilde{\phi}_0)^{1/2} \in W^{1,2}(D)$.

Proof. From Proposition 2.1, it follows that

$$0 < \inf_{x \in D} \frac{\tilde{\phi}_0(x)}{\phi_0(x)} \leqq \sup_{x \in D} \frac{\tilde{\phi}_0(x)}{\phi_0(x)} < \infty.$$

This is enough to show that in fact $(\phi_0\tilde{\phi}_0)^{1/2} \in C^1(\bar{D})$.

We now present a generalized Dirichlet or energy principle for ϕ_0. Actually, we obtain a family of mini-max variational principles—one for each k, where k is as above. In the selfadjoint case, that is, the case $a^{-1}b = \nabla Q$, our Dirichlet principle will reduce to the classical one if and only if $k = Q$ on $\partial D - \partial D_1$. (See Remark 3 after the statement of the theorem for a proof of this.) It does not seem possible to give just one mini-max variational principle, which reduces to the classical one simultaneously for every selfadjoint case.

THEOREM. *Let $L = L_0 - V \equiv \frac{1}{2}\nabla \cdot a\nabla + b \cdot \nabla - V$ on a bounded region $D \subset R^n$ with $C^{2,\alpha}$-boundary ∂D. Also let $\tilde{L} = \tilde{L}_0 - V \equiv \frac{1}{2}\nabla \cdot a\nabla - b \cdot \nabla - \nabla \cdot b - V$ be the adjoint to L. Assume that the $n \times n$ matrix a has entries $a_{ij} \in C^{1,\alpha}(\bar{D})$, that the n-vector b has components $b_i \in C^{1,\alpha}(\bar{D})$, that $V \in C^{\alpha}(\bar{D})$ and that f and k are in $C^{2,\alpha}(\bar{D})$ with $f \geqq 0$ and $f \not\equiv 0$ on ∂D. Define $\partial D_1 = \partial D \cap \{f = 0\}$. Assume $\operatorname{Re}(\sigma(L)) < 0$ where L is the above operator with the Dirichlet boundary condition on ∂D, and let ϕ_0 and $\tilde{\phi}_0$ be the unique solutions to $L\phi_0 = 0$ in D with $\phi_0 = f$ on ∂D and to $\tilde{L}\tilde{\phi} = 0$ in D with $\tilde{\phi}_0 = fe^{2k}$ on ∂D. Then the mini-max variational formula*

$$\lambda_0^{(k)} = \inf_{\substack{g = fe^k \text{ on } \partial D \\ g \in W^{1,2}(D) \\ (\operatorname{dist}(x, \partial D_1))^{-1} g(x) \in L^{\infty}(D)}} \sup_{\substack{h = k \text{ on } \partial D - \partial D_1 \\ h \in W^{1,2}(D, g^2\,dx)}} \left[\frac{1}{2} \int_D \left(\frac{\nabla g}{g} - a^{-1}b \right) a \left(\frac{\nabla g}{g} - a^{-1}b \right) g^2\,dx \right.$$

$$\left. - \frac{1}{2} \int_D (\nabla h - a^{-1}b) a (\nabla h - a^{-1}b) g^2\,dx + \int Vg^2\,dx \right] \tag{2.1}$$

is attained for a unique pair,

$$(g_0, h_0) \in \{(g, h) : g \in W^{1,2}(D), (\operatorname{dist}(x, \partial D_1))^{-1} g(x)) \in L^{\infty}(D), h \in W^{1,2}(D, g^2\,dx)\},$$

and in fact $g_0 = (\phi_0\tilde{\phi}_0)^{1/2}$ and $h_0 = \frac{1}{2}\log \tilde{\phi}_0/\phi_0$. Thus

$$\lambda_0^{(k)} = \frac{1}{2} \int_D \left(\frac{1}{2} \frac{\nabla\tilde{\phi}_0}{\tilde{\phi}_0} + \frac{1}{2} \frac{\nabla\phi_0}{\phi_0} - a^{-1}b \right) a \left(\frac{1}{2} \frac{\nabla\tilde{\phi}_0}{\tilde{\phi}_0} + \frac{1}{2} \frac{\nabla\phi_0}{\phi_0} - a^{-1}b \right) \phi_0\tilde{\phi}_0\,dx$$

$$- \frac{1}{2} \int_D \left(\frac{1}{2} \frac{\nabla\tilde{\phi}_0}{\tilde{\phi}_0} - \frac{1}{2} \frac{\nabla\phi_0}{\phi_0} - a^{-1}b \right) a \left(\frac{1}{2} \frac{\nabla\tilde{\phi}_0}{\tilde{\phi}_0} - \frac{1}{2} \frac{\nabla\phi_0}{\phi_0} - a^{-1}b \right) \phi_0\tilde{\phi}_0\,dx \tag{2.2a}$$

$$+ \int_D V\phi_0\tilde{\phi}_0\,dx.$$

In the special case $\{f = 0 \text{ or } f = 1\} \cap \partial D = \partial D$, then

$$\lambda_0^{(k)} = \int_D \left(\frac{1}{2} \frac{\nabla\phi_0 a \nabla\phi_0}{\phi_0^2} + V \right) \phi_0\tilde{\phi}_0\,dx. \tag{2.2b}$$

Formula (2.2b) *also holds for arbitrary* f *in the selfadjoint case* $a^{-1}b = \nabla Q$ *if* $Q = k$ *on* $\partial D - \partial D_1$.

Furthermore, the mini-max principle in (2.1) *may be converted to the following minimum principle*:

$$\mu_0^{(k)} = \inf_{\substack{g = fe^k \text{ on } \partial D \\ g \in W^{1,2}(D) \\ (\operatorname{dist}(x, \partial D_1))^{-1} g(x) \in L^\infty(D)}} \inf_{\substack{z = (z_1, \cdots, z_n) \in C^1(D) \\ \nabla \cdot (g^2 z) = 0 \text{ in } D}} \left[\frac{1}{2} \int_D \left(\frac{\nabla g}{g} - a^{-1} b \right) a \left(\frac{\nabla g}{g} - a^{-1} b \right) g^2 \, dx \right. \tag{2.3}$$

$$\left. + \int_D \left(\frac{1}{2}(zaz) + z(b - a \nabla k) \right) g^2 \, dx \right]$$

is attained uniquely at the pair (g_0, z_0), *where* $z_0 = \nabla h_0 - a^{-1} b$ *and* g_0 *and* h_0 *are as defined above. Moreover,* $\mu_0^{(k)} = \lambda_0^{(k)}$.

Before proving the theorem, we make a number of comments.

Remark 1. The theorem ought to hold without the condition $(\operatorname{dist}(x, \partial D_1))^{-1} g(x) \in L^\infty(D)$. Unfortunately, our proof requires this technical condition.

Remark 2. The smoothness conditions on a, b, V, D, f and k may be ignored as long as it is known that (possibly weak) solutions $\phi_0 > 0$ and $\tilde{\phi}_0 > 0$ exist in $W^{1,2}(D)$ and that $(\phi_0 \tilde{\phi}_0)^{1/2}$ is in $W^{1,2}(D)$. The only change now is that the condition $(\operatorname{dist}(x, \partial D_1))^{-1} g(x) \in L^\infty(D)$ should be replaced by $\phi_0^{-1} g \in L^\infty(D)$. It is true that in the proof, we integrate by parts and use the assumption that ϕ_0 and $\tilde{\phi}_0$ are C^2-functions. However, this is merely a convenience.

Remark 3. In the selfadjoint case, we have $a^{-1} b = \nabla Q$. This implies that $Q \in C^{2,\alpha}(\bar{D})$. Thus, it is always possible to pick k such that $k = Q$ on $\partial D - \partial D_1$ (in fact one might as well pick $k \equiv Q$). We noted, prior to stating the theorem, that only in the case $k = Q$ on $\partial D - \partial D_1$ will the variational principle reduce to the classical one given in (1.2). To see this note that only in this case will

$$\inf_{\substack{h = k \text{ on } \partial D - \partial D_1 \\ h \in W^{1,2}(D, g^2 dx)}} \left[\int_D (\nabla h - \nabla Q) a (\nabla h - \nabla Q) g^2 \, dx \right] = 0 \quad \text{for all } g \in W^{1,2}(D).$$

If this above expression is zero for all $g \in W^{1,2}(D)$, then it is easy to see that the condition $(\operatorname{dist}(x, \partial D_1))^{-1} g(x) \in L^\infty(D)$ is no longer needed. Thus (2.1) reduces to

$$\lambda_0^{(k)} = \inf_{\substack{g \in fe^k \text{ on } \partial D \\ g \in W^{1,2}(D)}} \left[\frac{1}{2} \int_D \left(\frac{\nabla g}{g} - \nabla Q \right) a \left(\frac{\nabla g}{g} - \nabla Q \right) g^2 \, dx + \int_D V g^2 \, dx \right]$$

$$= \inf_{\substack{u = f \text{ on } \partial D \\ u \in W^{1,2}(D)}} \left[\frac{1}{2} \int (\nabla u a \nabla u) \, e^{2Q} \, dx + \int_D V u^2 \, e^{2Q} \, dx \right],$$

where we have made the substitution $g = ue^Q$. This is (1.2). Furthermore (2.2b) now becomes

$$\lambda_0^{(k)} = \frac{1}{2} \int_D (\nabla \phi_0 a \nabla \phi_0) \, e^{2Q} \, dx + \int_D V \phi_0^2 \, e^{2Q} \, dx,$$

since $h_0 = Q$ and $\tilde{\phi}_0 / \phi_0 = e^{2Q}$.

Remark 4. Since ∂D may be composed of two disconnected pieces (D an annulus, for example), the special case $\{f = 0 \text{ or } f = 1\} \cap \partial D = \partial D$ does not only include the

trivial cases $f \equiv 0$ on ∂D and $f \equiv 1$ on ∂D. We note the reduction of (2.2a) to (2.2b) for two reasons. First of all, if $V \geqq 0$, then excluding the selfadjoint case $a^{-1}b = \nabla Q$ for some Q satisfying $Q = k$ on ∂D (see Remark 3), it is clear from (2.1) that only in the case $\{f = 0 \text{ or } f = 1\} \cap \partial D = \partial D$ can one guarantee by inspection that $\lambda_0^{(k)} \geqq 0$. This is reflected in (2.2a) and (2.2b). Also, in a probabilistic application of this theorem, which appears in [6], we need the form appearing in (2.2b).

Remark 5. This generalized Dirichlet principle is less general than the classical one since it requires that the boundary values be nonnegative. Indeed, if f changes sign, then ϕ_0 and $\tilde{\phi}_0$ may change sign and, consequently, neither g_0 nor h_0 is defined. However, this restriction is not too severe. For example, if $V = 0$ and if $\hat{\phi}_0$ satisfies $L\hat{\phi}_0 = 0$ on D and $\hat{\phi}_0 = \hat{f}$ on ∂D, for a general $\hat{f}$, then for sufficiently large M, $\phi_0 \equiv \hat{\phi}_0 + M$ satisfies $L\phi_0 = 0$ on D with $\phi_0 = f \equiv \hat{f} + M \geqq 0$ on ∂D. Thus $\hat{\phi}_0$ may be described as $\phi_0 - M$ where ϕ_0 has a generalized Dirichlet principle associated with it.

Proof. For $g \in W^{1,2}(D)$ satisfying $g = fe^k$ on ∂D, let

$$H_g(h) = \frac{1}{2} \int_D (\nabla h - a^{-1}b) a (\nabla h - a^{-1}b) g^2 \, dx.$$

First we show that

$$\inf_{\substack{h = k \text{ on } \partial D - \partial D_1 \\ h \in W^{1,2}(D, g^2\,dx)}} H_g(h) \quad \text{is attained at a unique } h = h_g \in W^{1,2}(D, g^2\,dx).$$

Actually, the proof is almost identical to a special case of the proof of a result which can be found in [5, pp. 688-690]. For completeness, we will give a proof although we will not write down all the details. By the Schwarz inequality, it is easy to obtain $H_g(h) \geqq c_1 \int_D |\nabla h|^2 g^2 \, dx - c_2$ for positive constants c_1 and c_2. Thus if $\{h_n\}$ is a minimizing sequence, then $\int_D |\nabla h_n|^2 g^2 \, dx$ is bounded independent of n. This, coupled with the fact that $h_n = k$ on $\partial D - \partial D_1$ gives weak compactness in $W^{1,2}(D, g^2\,dx)$. Now $H_g(h)$ is lower semicontinuous under weak convergence since the norm is lower semicontinuous under weak convergence. Thus, in fact, any limit $\hat{h}$ of a subsequence $\{h_n\}$ must be a minimizer. Varying $H_g(h)$ at $\hat{h}$ gives

$$\int_D (\nabla \hat{h} - a^{-1}b) a \nabla q g^2 \, dx = 0, \tag{2.4}$$

for all $q \in W^{1,2}(D, g^2\,dx)$ satisfying $q = 0$ on $\partial D - \partial D_1$. Now if $\tilde{h}$ is also a minimizer, then (2.4) also holds with $\tilde{h}$ in place of $\hat{h}$. Subtracting gives $\int_D (\nabla \hat{h} - \nabla \tilde{h}) a \nabla q g^2 \, dx = 0$. Substituting $q = \hat{h} - \tilde{h}$ shows that $\nabla \hat{h} - \nabla \tilde{h} = 0$ a.e. $[g^2\,dx]$, and since $\hat{h} = \tilde{h} = k$ on $\partial D - \partial D_1$, we have $\hat{h} = \tilde{h}$ a.e. $[g^2\,dx]$. This completes the proof. We will denote the minimizer by h_g.

Consider $g = g_0 \equiv (\phi_0 \tilde{\phi}_0)^{1/2}$. By Proposition 2.2, $g_0 \in W^{1,2}(D)$. We claim that $h_0 \equiv h_{g_0} = \frac{1}{2} \log (\tilde{\phi}_0 / \phi_0)$. Note that h_0 satisfies the appropriate boundary conditions since $\phi_0 = f$ and $\tilde{\phi}_0 = fe^{2k}$ on ∂D. We must check that (2.4) is satisfied with $\hat{h}$ replaced by h_0. From the fact that $L\phi_0 = 0$ and $\tilde{L}\tilde{\phi}_0 = 0$, one can check that h_0 satisfies

$$\nabla \cdot a \nabla h_0 + 2 \frac{\nabla g_0}{g_0} \nabla h_0 = \nabla \cdot b + 2 \frac{\nabla g_0}{g_0} b. \tag{2.5}$$

Integrating the left-hand side of (2.4) by parts and using (2.5), one sees that (2.4) is indeed satisfied with $\hat{h} = h_0$.

Now, for $g \in W^{1,2}(D)$, let $\gamma = g^2$ and define for γ and for $h \in W^{1,2}(D, \gamma\,dx)$, $\psi(h, \gamma) = \frac{1}{2}\int_D (\nabla h - a^{-1}b)a(\nabla h - a^{-1}b)\gamma\,dx$. Also define

$$J(\gamma) = \psi(\tfrac{1}{2}\log\gamma, \gamma) - \inf_{\substack{h=k \text{ on } \partial D - \partial D_1 \\ h \in W^{1,2}(D, \gamma\,dx)}} \psi(h, \gamma) + \int_D V\gamma\,dx.$$

Then the right-hand side of (2.1) may be written as

$$\inf_{\substack{\gamma = f^2 e^{2k} \text{ on } \partial D \\ \gamma^{1/2} \in W^{1,2}(D) \\ (\operatorname{dist}(x, \partial D_1))^{-1}\gamma^{1/2}(x) \in L^\infty(D)}} J(\gamma).$$

To show that (a) the mini-max in (2.1) is attained at some pair (g, h_g); (b) the pair is unique; and (c) the pair is given by (g_0, h_0), it obviously suffices to show that $J(\gamma_0) < J(\gamma_1)$ for all $\gamma_1 \neq \gamma_0 \equiv g_0^2$ that satisfy $\gamma_1 = f^2 e^{2k}$ on ∂D, $\gamma_1^{1/2} \in W^{1,2}(D)$ and $(\operatorname{dist}(x, \partial D_1))^{-1}\gamma_1^{1/2}(x) \in L^\infty(D)$. To do this, we will show that $M(p) \equiv J(\gamma_p)$ is strictly convex on $[0, 1]$ and satisfies $M'(0) = 0$, where $\gamma_p = p\gamma_1 + (1-p)\gamma_0$.

Now

$$M(p) = J(\gamma_p) = \psi\left(\frac{1}{2}\log\gamma_p, \gamma_p\right) + \left(-\inf_{\substack{h=k \text{ on } \partial D - \partial D_1 \\ h \in W^{1,2}(D, \gamma_p dx)}} \psi(h, \gamma_p)\right) + \int_D V\gamma_p\,dx$$

and it is easy to see that the second term on the right-hand side is convex in p (the p appearing in the domain over which the infimum is calculated causes no problem). Since the third term on the right-hand side is linear, to show strict convexity we need only show that $(d^2/dp^2)(\psi(\frac{1}{2}\log\gamma_p, \gamma_p)) > 0$ for $p \in [0, 1]$. We have

$$\psi\left(\frac{1}{2}\log\gamma_p, \gamma_p\right) = \frac{1}{2}\int_D \frac{(\frac{1}{2}\nabla\gamma_p - a^{-1}b\gamma_p)a(\frac{1}{2}\nabla\gamma_p - a^{-1}b\gamma_p)}{\gamma_p}\,dx$$

and

$$\frac{d^2}{dp^2}\left(\psi\left(\frac{1}{2}\log\gamma_p, \gamma_p\right)\right) = \int_D \left(\frac{\nabla(\gamma_1 - \gamma_0) - a^{-1}b(\gamma_1 - \gamma_0)}{\gamma_p^{1/2}} - \frac{\nabla\gamma_p - a^{-1}b\gamma_p(\gamma_1 - \gamma_0)}{\gamma_p^{3/2}}\right)$$
$$\cdot a\left(\frac{\nabla(\gamma_1 - \gamma_0) - a^{-1}b(\gamma_1 - \gamma_0)}{\gamma_p^{1/2}} - \frac{\nabla\gamma_p - a^{-1}b\gamma_p(\gamma_1 - \gamma_0)}{\gamma_p^{3/2}}\right) dx \geqq 0.$$

Equality occurs if and only if

$$\frac{\nabla(\gamma_1 - \gamma_0) - a^{-1}b(\gamma_1 - \gamma_0)}{\gamma_p^{1/2}} = \frac{\nabla\gamma_p - a^{-1}b\gamma_p(\gamma_1 - \gamma_0)}{\gamma_p^{3/2}},$$

that is, only if $(\nabla(\gamma_1 - \gamma_0))/\gamma_1 - \gamma_0 = \nabla p/\gamma_p$. For this to occur for even one fixed p requires that γ_1 be a multiple of γ_0, which is impossible since by assumption $\gamma_1 \not\equiv \gamma_0$ and $\gamma_1 = \gamma_0 = f^2 e^{2k}$ on ∂D. This proves the strict convexity.

We now prove that $M'(0) = 0$. Let $h_p \in W^{1,2}(D, \gamma_p)$ be the function at which

$$\inf_{\substack{h=k \text{ on } \partial D - \partial D_1 \\ h \in W^{1,2}(D, \gamma_p)}} \int_D (\nabla h - a^{-1}b)a(\nabla q - a^{-1}b)\gamma_p\,dx$$

is attained. From (2.4), h_p satisfies

$$\int (\nabla h_p - a^{-1}b)a\nabla q\gamma_p\,dx = 0, \tag{2.6}$$

for all $q \in W^{1,2}(D, \gamma_p\, dx)$ satisfying $q = 0$ on $\partial D - \partial D_1$. We will show that

$$\nabla h_0' \equiv \lim_{\varepsilon \to 0} \frac{\nabla h_\varepsilon - \nabla h_0}{\varepsilon} \tag{2.7}$$

exists as a weak limit in $L^2(D, \gamma_0\, dx)$. Then we have

$$\begin{aligned} M(p) = &\frac{1}{2}\int_D \frac{(\tfrac{1}{2}\nabla\gamma_p - a^{-1}b\gamma_p)a(\tfrac{1}{2}\nabla\gamma_p - a^{-1}b\gamma_p)}{\gamma_p}\, dx \\ &-\frac{1}{2}\int_D (\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b)\gamma_p\, dx + \int_D V\gamma_p\, dx \end{aligned} \tag{2.8}$$

and, formally,

$$\begin{aligned} M'(0) = &\int_D \left[\frac{(\tfrac{1}{2}\nabla(\gamma_1 - \gamma_0) - a^{-1}b(\gamma_1 - \gamma_0))a(\tfrac{1}{2}\nabla\gamma_0 - a^{-1}b\gamma_0)}{\gamma_0}\right. \\ &\left. -\frac{(\tfrac{1}{2}\nabla\gamma_0 - a^{-1}b\gamma_0)a(\tfrac{1}{2}\nabla\gamma_0 - a^{-1}b\gamma_0)(\gamma_1 - \gamma_0)}{2\gamma_0^2}\right] dx \\ &-\int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0')\gamma_0\, dx \\ &-\frac{1}{2}\int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0 - a^{-1}b)(\gamma_1 - \gamma_0)\, dx + \int_D V(\gamma_1 - \gamma_0)\, dx. \end{aligned} \tag{2.9}$$

Under the assumption that (2.7) and (2.9) hold rigorously, we will show that $M'(0) = 0$. Then we will go back and prove (2.7) and (2.9). The term $\int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0')\gamma_0\, dx = 0$ by (2.6) and (2.7). Thus (2.9) becomes

$$\begin{aligned} M'(0) = &\int_D \left[\frac{(\tfrac{1}{2}\nabla(\gamma_1 - \gamma_0) - a^{-1}b(\gamma_1 - \gamma_0))a(\tfrac{1}{2}\nabla\gamma_0 - a^{-1}b\gamma_0)}{\gamma_0}\right. \\ &\left. -\frac{(\tfrac{1}{2}\nabla\gamma_0 - a^{-1}b\gamma_0)a(\tfrac{1}{2}\nabla\gamma_0 - a^{-1}b\gamma_0)(\gamma_1 - \gamma_0)}{2\gamma_0^2}\right] dx \\ &-\frac{1}{2}\int_D (\nabla h - a^{-1}b)a(\nabla h_0 - a^{-1}b)(\gamma_1 - \gamma_0)\, dx \\ &+\int_D V(\gamma_1 - \gamma_0)\, dx. \end{aligned} \tag{2.10}$$

From the fact that $\phi_0 = g_0\, e^{-h_0}$ and $\tilde{\phi}_0 = g_0\, e^{h_0}$, one can verify that $\gamma_0 = g_0^2$ satisfies

$$\frac{1}{2}\nabla \cdot a\nabla\gamma_0 - \frac{1}{4}\frac{\nabla\gamma_0 a\nabla\gamma_0}{\gamma_0} + \gamma_0(\nabla h_0 a\nabla h_0 - \nabla \cdot b - 2\nabla h_0 b) - V\gamma_0 = 0. \tag{2.11}$$

Now (2.9) arises as the variation of (2.8) in the direction $\gamma_1 - \gamma_0$. If we set $q = \gamma_1 - \gamma_0$ in (2.9), and then integrate by parts and use (2.11) and the fact that $q = 0$ on ∂D, we obtain $M'(0) = 0$.

It remains to show (2.7) and (2.9). First we show (2.7). Since $\int_D (\nabla h_p - b)a(\nabla h_p - b)\gamma_p\, dx$ is bounded independent of p, it is clear that

$$\overline{\lim_{p \to 0}} \int_D |\nabla h_p|^2 \gamma_0\, dx < \infty. \tag{2.12}$$

By the assumption that $(\operatorname{dist}(x, \partial D_1))^{-1}\gamma_1^{1/2}(x) \in L^\infty(D)$ and by Proposition 2.1, (2.12) guarantees that

$$\overline{\lim_{p \to 0}} \int_D |\nabla h_p|^2 \gamma_1\, dx < \infty. \tag{2.13}$$

From (2.6), we have $\int_D (\nabla h_p - a^{-1}b)a\nabla q\gamma_p\,dx = 0$ for $0<p<1$ and $\int(\nabla h_0 - a^{-1}b)a\nabla q\gamma_0\,dx = 0$ for all $q \in W^{1,2}(D, \gamma_0\,dx)$ satisfying $q=0$ on $\partial D - \partial D_1$. Subtracting, we obtain

$$\int_D (\nabla h_p - \nabla h_0)a\nabla q\gamma_0\,dx + p\int_D (\nabla h_p - b)a\nabla q(\gamma_1-\gamma_0)\,dx = 0. \tag{2.14}$$

From (2.12), (2.13) and (2.14), we have

$$\lim_{p\to 0}\int_D (\nabla h_p - \nabla h_0)a\nabla q\gamma_0\,dx = 0, \tag{2.15}$$

for all $q \in W^{1,2}(D, \gamma_0\,dx)$ satisfying $q=0$ on $\partial D - \partial D_1$. Since $h_p - h_0$ also satisfies $h_p - h_0 \in W^{1,2}(D, \gamma_0\,dx)$ and $h_p - h_0 = 0$ on $\partial D - \partial D_1$, (2.15) guarantees that $\nabla h_p \to \nabla h_0$ weakly in $L^2(D, \gamma_0\,dx)$. Using this fact, we see from (2.14) that

$$\lim_{p\to 0}\int_D \left(\frac{\nabla h_p - \nabla h_0}{p}\right) a\nabla q\gamma_0\,dx = -\int_D (\nabla h_0 - b)a\nabla q(\gamma_1-\gamma_0)\,dx.$$

This shows that $(\nabla h_p - \nabla h_0)/p \to (\nabla h_0 - b)(\gamma_0-\gamma_1)/\gamma_0 \equiv \nabla h_0'$ weakly in $L^2(D, \gamma_0\,dx)$ and proves (2.7).

To show (2.9), we must show that

$$\begin{aligned}
&\lim_{p\to 0}\frac{1}{p}\Bigg[\int (\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b)\gamma_p\,dx \\
&\qquad - \int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0 - a^{-1}b)\gamma_0\,dx\Bigg] \\
&= 2\int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0')\gamma_0\,dx \\
&\qquad + \int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0 - a^{-1}b)(\gamma_1-\gamma_0)\,dx.
\end{aligned} \tag{2.16}$$

We have

$$\begin{aligned}
&\frac{1}{p}\Bigg[\int_D (\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b)\gamma_p\,dx \\
&\qquad - \int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0 - a^{-1}b)\gamma_0\,dx\Bigg] \\
&= \frac{1}{p}\int_D \Big[(\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b) - (\nabla h_0 - a^{-1}b)a(\nabla h_0 - a^{-1}b)\Big]\gamma_0\,dx \\
&\qquad + \int_D (\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b)(\gamma_1-\gamma_0)\,dx \\
&= \frac{1}{p}\int_D ((\nabla h_p - a^{-1}b) + (\nabla h_0 - a^{-1}b))a((\nabla h_p - a^{-1}b) \\
&\qquad\qquad - (\nabla h_0 - a^{-1}b))\gamma_0\,dx \\
&\qquad + \int_D (\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b)(\gamma_1-\gamma_0)\,dx \\
&= \int_D ((\nabla h_p - a^{-1}b) + (\nabla h_0 - a^{-1}b))a\left(\frac{\nabla h_p - \nabla h_0}{p}\right)\gamma_0\,dx
\end{aligned} \tag{2.17}$$

$$+\int_D (\nabla h_p - a^{-1}b)a(\nabla h_p - a^{-1}b)(\gamma_1 - \gamma_0)\,dx.$$

Since $(\nabla h_p - \nabla h_0)/p$ converges weakly in $L^2(D, \gamma_0\,dx)$, $\nabla h_p \to \nabla h_0$ strongly in $L^2(D, \gamma_0\,dx)$. By Proposition 2.1 and the assumption $(\text{dist}\,(x, \partial D_1))^{-1}\gamma_1^{1/2}(x) \in L^\infty(D)$, we also have $\nabla h_p \to \nabla h_0$ strongly in $L^2(D, \gamma_1\,dx)$. The weak convergence of $(\nabla h_p - \nabla h_0)/p$ to $\nabla h_0'$ in $L^2(D, \gamma_0\,dx)$ and the strong convergence of ∇h_p to ∇h_0 in $L^2(D, \gamma_0\,dx)$ and in $L^2(D, \gamma_1\,dx)$ show that as $p \to 0$, the right-hand side of (2.17) converges to the right-hand side of (2.16). This verifies (2.16).

Now we derive (2.2a) and (2.2b). We obtain (2.2a) simply by substituting $g_0 = (\phi_0\tilde{\phi}_0)^{1/2}$ and $h_0 = \frac{1}{2}\log(\tilde{\phi}_0/\phi_0)$ in (2.1). For the case $a^{-1}b = \nabla Q$ and $Q = k$ on ∂D, (2.2b) was proven in Remark 3. We now show that (2.2b) holds in the case $\{f = 0$ or $f = 1\} \cap \partial D = \partial D$. By (2.4), we have

$$\int_D (\nabla h_0 - a^{-1}b)a\nabla q g^2\,dx = 0,$$

for all $q \in W^{1,2}(D, g^2\,dx)$ satisfying $q = 0$ on $\partial D - \partial D_1$. Because of the condition on f, we may pick $q = \log g_0 - h_0$ to obtain

$$\int (\nabla h_0 - a^{-1}b)a\left(\frac{\nabla g_0}{g} - \nabla h_0\right)g_0^2\,dx = 0. \tag{2.18}$$

From (2.18) we have

$$\begin{aligned}\frac{1}{2}\int_D (\nabla h_0 - a^{-1}b)a(\nabla h_0 - a^{-1}b)g_0^2\,dx &= \int_D g_0\nabla g_0 a\nabla h_0\,dx - \int_D g_0 b\nabla g_0\,dx \\ &\quad -\frac{1}{2}\int_D (\nabla h_0 a\nabla h_0)g_0^2\,dx \\ &\quad +\frac{1}{2}\int_D (ba^{-1}b)g_0^2\,dx.\end{aligned}$$

Substituting this in (2.1), we obtain

$$\begin{aligned}&\frac{1}{2}\int_D \left(\frac{\nabla g_0}{g_0} - a^{-1}b\right)a\left(\frac{\nabla g_0}{g_0} - a^{-1}b\right)g_0^2\,dx - \int_D g_0\nabla g_0 a\nabla h_0\,dx + \int_D g_0 b\nabla g_0\,dx \\ &\qquad +\frac{1}{2}\int_D (\nabla h_0 a\nabla h_0)g_0^2\,dx - \frac{1}{2}\int_D (ba^{-1}b)g_0^2\,dx + \int_D Vg_0^2\,dx \\ &= \frac{1}{2}\int_D (\nabla g_0 - g_0\nabla h_0)a(\nabla g_0 - g_0\nabla h_0)\,dx + \int_D Vg_0^2\,dx \\ &= \int_D \left(\frac{1}{2}\frac{(\nabla\phi_0 a\nabla\phi_0)}{\phi_0^2} + V\right)\phi_0\tilde{\phi}_0\,dx,\end{aligned}$$

since

$$g_0 = (\phi_0\tilde{\phi}_0)^{1/2} \quad \text{and} \quad h_0 = \frac{1}{2}\log\frac{\tilde{\phi}_0}{\phi_0}.$$

We now turn to (2.3). Let $R_g(z) = \int_D (\frac{1}{2}zaz + z(b - a\nabla k))g^2\,dx$. Essentially the same proof as that used to prove the existence of a unique minimizer $h = h_g$ of $H_g(h)$ at the

beginning of the proof of the theorem shows that there exists a unique $\hat{z} = (\hat{z}_1, \cdots, \hat{z}_n) \in L^2(D, g^2\,dx)$ which is a weak solution of $\nabla \cdot (g^2 z) = 0$ in D and which satisfies

$$R_g(\hat{z}) = \inf_{\substack{z \in C^1(D) \\ \nabla \cdot (g^2 z) = 0 \text{ in } D}} R_g(z).$$

This proof also shows that $\hat{z}$ is the unique solution to

$$\int_D q(a\hat{z} + b - a\nabla k) g^2\,dx = 0, \tag{2.19}$$

for every $q \in L^2(D, g^2\,dx)$ which is a weak solution to $\nabla \cdot (g^2 q) = 0$ in D.

In order to complete the proof of (2.3), it is enough to show for each $g \in W^{1,2}(D)$ satisfying $g = fe^k$ on ∂D, that

(i) $\hat{z} = \nabla h_g - a^{-1}b$, where h_g minimizes $H_g(h)$ over $h \in W^{1,2}(D, g^2\,dx)$ satisfying $h = k$ on $\partial D - \partial D_1$, and that

(ii) $R_g(\hat{z}) = -\frac{1}{2}\int_D (\nabla h_g - a^{-1}b) a (\nabla h_g - a^{-1}b) g^2\,dx$.

To show that (i) holds, simply substitute $\hat{z} = \nabla h_g - a^{-1}b$ in (2.19), integrate by parts and use the condition on q and the boundary condition on h. Now (ii) follows by substituting $q = \hat{z} = \nabla h_g - a^{-1}b$ in (2.19).

REFERENCES

[1] M. D. DONSKER AND S. R. S. VARADHAN, *Second order elliptic differential operators*, Comm. Pure Appl. Math., 29 (1976), pp. 595-621.

[2] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.

[3] C. HOLLAND, *A minimum principle for the principle eigenvalue for second order linear elliptic equations with natural boundary conditions*, Comm. Pure Appl. Math., 31 (1978), pp. 509-519.

[4] M. G. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, AMS Translations, Ser. 1, 10 (1962), pp. 199-324.

[5] R. G. PINSKY, *The I-function for diffusion processes with boundaries*, Ann. Probab., 13 (1985), pp. 676-692.

[6] ———, *A mini-max variational formula giving necessary and sufficient conditions for recurrence or transience of multidimensional diffusion processes*, Ann. Probab., to appear.

[7] M. H. PROTTER AND H. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

016

ON THE DETERMINATION OF A FUNCTION FROM SPHERICAL AVERAGES*

LARS-ERIK ANDERSSON†

Abstract. Let f be a function $f:\mathbb{R}^{n+1}\to\mathbb{R}$ which is even in the last variable, i.e., such that $f(x,-y)=f(x,y)$ where $x\in\mathbb{R}^n$, $y\in\mathbb{R}$. The mapping R is defined by $f\mapsto Rf=g$ where $g(x,r)$ is the average of f over a sphere with radius r and center at a point $(x,0)$ in the hyperplane $y=0$. The problem to invert the mapping R is studied. Extending the domain of the mapping R to the class of tempered distributions, we give a characterization of the range of R and prove that the inverse mapping R^{-1} exists and is continuous in the topology of distributions. An inversion formula, first discovered by J. Fawcett, is obtained in terms of Fourier transforms and a Sobolev estimate for the inverse mapping is given. Next, inversion methods using only values of g on some bounded set are studied. First a uniqueness theorem of Courant and Hilbert is generalized to distributions. Inversion formulas involving partial Fourier transforms are given and a numerical inversion procedure is proposed.

Key words. inverse problem, spherical averages, generalized Radon transform, ill-posed

AMS(MOS) subject classification. 44A15

Introduction. Consider functions $f:\mathbb{R}^{n+1}\to\mathbb{R}$ which are even in the last variable, i.e., such that $f(x,-y)=f(x,y)$ where $x\in\mathbb{R}^n$, $y\in\mathbb{R}$. Define the mapping $f\mapsto Rf=g$ by letting $g(x,r)$ be the average of f over a sphere with radius r and center at the point $(x,0)$ in the hyperplane $y=0$. Consequently,

$$g(x,r)=\frac{1}{\omega_n}\int_{S^n} f(x+r\xi,r\eta)\,dS_n(\xi,\eta),$$

where S^n is the unit sphere $\{(\xi,\eta):|\xi|^2+\eta^2=1\}$ in $\mathbb{R}^{n+1}$, dS_n is the surface measure on a sphere and ω_n is the area of S^n. The domain of this mapping R will be specified later.

The problem of inverting the mapping R, i.e., to determine the function f when g is known, which is the subject of this paper, has been treated by several authors. The uniqueness of the solution f in the class of continuous functions was proved by Courant and Hilbert [2]. Employing essentially the same technique as Courant and Hilbert, Lavrentiev et al. [9] have given inversion methods as well as a characterization of the range of R. Unfortunately, these methods are difficult to use numerically. In a recent paper Fawcett [3] has given an inversion method using Fourier–Hankel transformation and a so-called back projection operator. He thus provides a method to solve the global problem, i.e., to determine $f(x,y)$ for all $(x,y)\in\mathbb{R}^{n+1}$ when $g(x,r)$ is known for all $x\in\mathbb{R}^n$, $r>0$.

The inversion problem for the mapping R, belongs to the same class of problems as the inversion problems for the classical Radon transform and its generalizations (see for instance Helgason [4]). The problem that is studied in the present paper is of interest in several applications, for example in image processing of so-called synthetic aperture radar (SAR) data, when using wavelengths of the magnitude 3–30 m. In such a situation, when the wavelength is considerably larger than the dimension of the antenna, the emitted radar signal is spread more or less uniformly in all directions, and then the problem of inverting spherical averages enters naturally in the analysis. For a more detailed account of this application we refer to [5]–[7].

* Received by the editors December 18, 1985; accepted for publication (in revised form) February 13, 1987.

† Department of Mathematics, Linköping Institute of Technology, S-581 83 Linköping, Sweden.

Other interesting applications are found in connection with inverse problems for hyperbolic partial differential equations. One example is the linearized inverse scattering problem in acoustics as described by Cohen and Bleistein [1]. Surveys of these applications are given by Lavrentiev et al. [9] and by Romanov [11].

One of the main results of the present work, given in § 2, is a rigorous derivation of Fawcett's inversion formula under very general conditions on f and g. Our (slightly different) version of the formula is given completely in terms of Fourier transforms. In § 4 we turn instead to the local inversion problem, i.e., how to determine f when $g(x, r)$ is known in some region $|x| < A \leqq \infty$, $0 < r < B < \infty$. Using the theory of § 2, we first deduce a theorem of uniqueness which is a generalization of Courant's to the case when f is allowed to be a distribution. In § 5 we give inversion formulas for the local problem in terms of integral equations involving partial Fourier transforms of f and g. We also discuss the range of the mapping R in the local case and the ill-posedness of the problem, and propose an inversion procedure.

1. Some notation and definitions. By $\mathcal{S}(\mathbb{R}^n)$ we mean the so-called Schwartz class of infinitely differentiable functions φ for which the norms $\|\varphi\|_N = \sup\{|x|^k |D^\alpha \varphi(x)| : 0 \leqq k \leqq N,\ 0 \leqq |\alpha| \leqq N,\ x \in \mathbb{R}^n\}$ are finite for all $N \geqq 0$. Here $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ is a multiindex. $\mathcal{S}(\mathbb{R}^n)$ is given the topology which is induced by these norms.

$$\mathcal{S}_e(\mathbb{R}^{n+1}) = \{\psi \in \mathcal{S}(\mathbb{R}^{n+1}) : \psi(x, -y) = \psi(x, y),\ \forall x \in \mathbb{R}^n,\ y \in \mathbb{R}\}.$$

For convenience we will in this paper consider the functions $g(x, r)$ where $x \in \mathbb{R}^n$, $r \geqq 0$, to be defined on $\mathbb{R}^n \times \mathbb{R}^{n+1}$ and to depend radially on the last $n+1$ variables. We therefore introduce

$$\mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1}) = \{\varphi \in \mathcal{S}(\mathbb{R}^{2n+1}) : \varphi(x, z) = \varphi(x, Uz),\ \forall x \in \mathbb{R}^n,\ z \in \mathbb{R}^{n+1}$$
$$\text{and for all orthonormal transformations } U\}.$$

$\mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ is thus the subspace of functions in $\mathcal{S}(\mathbb{R}^{2n+1})$ which depend radially on the last $n+1$ variables. $\mathcal{S}_e(\mathbb{R}^{n+1})$ and $\mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ are given the same topology as $\mathcal{S}(\mathbb{R}^{n+1})$ and $\mathcal{S}(\mathbb{R}^{2n+1})$, respectively. $\mathcal{S}'(\mathbb{R}^n)$, the class of tempered distributions on $\mathbb{R}^n$, is by definition the set of all continuous linear functionals on $\mathcal{S}(\mathbb{R}^n)$. $\mathcal{S}'(\mathbb{R}^n)$ is given the weak topology. $\langle f, \psi\rangle$ denotes the value of $f \in \mathcal{S}'(\mathbb{R}^n)$ at $\psi \in \mathcal{S}(\mathbb{R}^n)$. $\mathcal{S}'_e(\mathbb{R}^{n+1}) = \{f \in \mathcal{S}'(\mathbb{R}^{n+1}) : \langle f, \psi\rangle = 0$ whenever $\psi(x, -y) = -\psi(x, y)$ for all $x, y\}$, i.e., $\mathcal{S}'_e(\mathbb{R}^{n+1})$ is the subspace of tempered distributions that are even in the last variable. $\mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1}) = \{g \in \mathcal{S}'(\mathbb{R}^{2n+1}) : \langle g, \varphi\rangle = 0$ whenever $\int_{|z|=r} \varphi(x, z)\, dS_n(z) = 0$ for all $x \in \mathbb{R}^n$ and $r > 0\}$, i.e., $\mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ is the subspace of tempered distributions which depend radially on the last $n+1$ variables. $\mathcal{S}'_e(\mathbb{R}^{n+1})$ and $\mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ may also be identified with the class of continuous linear functionals on $\mathcal{S}_e(\mathbb{R}^{n+1})$ and $\mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ respectively. In the following we will occasionally write $\mathcal{S}$, $\mathcal{S}_e$, $\mathcal{S}_r$, etc., instead of $\mathcal{S}(\mathbb{R}^n)$, $\mathcal{S}_e(\mathbb{R}^{n+1})$, $\mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$. Fourier transformation is defined by $\mathcal{F}\varphi(\sigma) = \hat{\varphi}(\sigma) = \int_{\mathbb{R}^n} e^{-i\langle\sigma, x\rangle} \varphi(x)\, dx$, $\mathcal{F}^{-1}\hat{\varphi}(x) = \varphi(x) = (1/(2\pi)^n) \int_{\mathbb{R}^n} e^{i\langle\sigma, x\rangle} \hat{\varphi}(\sigma)\, d\sigma$ when $\varphi \in \mathcal{S}(\mathbb{R}^n)$ and by $\langle -\hat{f}, \hat{\varphi}\rangle = (2\pi)^n \langle f, \varphi\rangle$ when $f \in \mathcal{S}'(\mathbb{R}^n)$.

We note that Fourier transformation gives isomorphisms

$$\mathcal{F}: \mathcal{S} \to \mathcal{S}, \qquad \mathcal{F}: \mathcal{S}' \to \mathcal{S}',$$
$$\mathcal{F}: \mathcal{S}_e \to \mathcal{S}_e, \qquad \mathcal{F}: \mathcal{S}'_e \to \mathcal{S}'_e,$$
$$\mathcal{F}: \mathcal{S}_r \to \mathcal{S}_r, \qquad \mathcal{F}: \mathcal{S}'_r \to \mathcal{S}'_r.$$

If $\varphi, \psi \in \mathcal{S}(\mathbb{R}^n)$ then $\varphi * \psi(x) = \int_{\mathbb{R}^n} \varphi(x - z)\psi(z)\, dz$ is the convolution. $(\varphi * \psi)\hat{} = \hat{\varphi}\hat{\psi}$. If $f \in \mathcal{S}'(\mathbb{R}^n)$ then $f * \varphi$ is defined by $\langle f * \varphi, \psi\rangle = \langle f, \check{\bar{\varphi}} * \psi\rangle$. Here $\check{\varphi}$ is defined by

$\check{\varphi}(x)=\varphi(-x)$ and $\bar{\varphi}$ denotes the complex conjugate of φ. Elements of $\mathcal{S}_e(\mathbb{R}^{n+1})$ or $\mathcal{S}'_e(\mathbb{R}^{n+1})$ will be denoted by, for example, $\psi(x, y)$ or $f(x, y)$ and their Fourier transforms by $\hat{\psi}(\sigma, \omega)$ or $\hat{f}(\sigma, \omega)$. Here $x, \sigma \in \mathbb{R}^n$ and $y, \omega \in \mathbb{R}$.

Elements of $\mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ or $\mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ are written, for example, $\varphi(x, r)$ or $g(x, r)$ and their Fourier transforms $\hat{\varphi}(\sigma, \rho)$ or $\hat{g}(\sigma, \rho)$. Here $x, \sigma \in \mathbb{R}^n$ and $r, \rho \in \mathbb{R}$, r and $\rho \geqq 0$. In § 5 we use the notation ψ^*, φ^*, etc. for the partial Fourier transform with respect to the n first variables.

$$\psi^*(\sigma, y)=\int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\psi(x, y)\,dx, \qquad \varphi^*(\sigma, r)=\int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\varphi(x, r)\,dx$$

when $\psi \in \mathcal{S}_e$ and $\varphi \in \mathcal{S}_r$.

$$\langle f^*, \psi^*\rangle=(2\pi)^n\langle f, \psi\rangle \quad \text{and} \quad \langle g^*, \varphi^*\rangle=(2\pi)^n\langle g, \varphi\rangle$$

when $f \in \mathcal{S}'_e$ and $g \in \mathcal{S}'_r$.

Note that $\psi \in \mathcal{S}_e \Leftrightarrow \psi^* \in \mathcal{S}_e$, $f \in \mathcal{S}'_e \Leftrightarrow f^* \in \mathcal{S}'_e$, $\varphi \in \mathcal{S}_r \Leftrightarrow \varphi^* \in \mathcal{S}_r$ and $g \in \mathcal{S}'_r \Leftrightarrow g^* \in \mathcal{S}'_r$. In § 5 we also use J_ν and I_ν to denote respectively the ordinary and the modified Bessel functions of the first kind and of order ν. See Magnus and Oberhettinger [10]. $(\mathcal{L}f)(p)=\tilde{f}(p)=\int_0^\infty e^{-pt}f(t)\,dt$ denotes the Laplace transform of the function f. Finally $[x]$ denotes the integer part of x.

2. An inversion formula for the global problem. It is almost obvious that the relation

$$g(x, r)=\frac{1}{\omega_n}\int_{S^n} f(x+r\xi, r\eta)\,dS_n(\xi, \eta)=Rf(x, r)$$

defines a linear, continuous mapping

$$R: \mathcal{S}_e(\mathbb{R}^{n+1}) \to \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1}).$$

Note, however, that in general, $Rf \notin L^1(\mathbb{R}^{2n+1})$ or $L^2(\mathbb{R}^{2n+1})$ when $f \in \mathcal{S}_e(\mathbb{R}^{n+1})$, not even if f has compact support. Now $\mathcal{S}_e(\mathbb{R}^{n+1})$ may also be thought of as a linear subspace of $\mathcal{S}'_e(\mathbb{R}^{n+1})$, equipped with the topology of $\mathcal{S}'_e(\mathbb{R}^{n+1})$. It is well known, then, that $\mathcal{S}_e(\mathbb{R}^{n+1})$ is dense in $\mathcal{S}'_e(\mathbb{R}^{n+1})$.

THEOREM 2.1. *If $\mathcal{S}_e(\mathbb{R}^{n+1})$ is given the topology of $\mathcal{S}'_e(\mathbb{R}^{n+1})$ then the mapping*

$$R: \mathcal{S}_e(\mathbb{R}^{n+1}) \to \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$$

is continuous and can, by continuity, be extended to a mapping (with abuse of notation)

$$R: \mathcal{S}'_e(\mathbb{R}^{n+1}) \to \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1}).$$

Further, the range of this extended mapping R is the closed subspace

$$\mathcal{S}'_{r,\mathrm{cone}}(\mathbb{R}^n \times \mathbb{R}^{n+1})=\{g \in \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1}): \operatorname{supp} \hat{g} \subset \{(\sigma, \rho): \rho \geqq |\sigma|\}\}$$
$$\subset \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1}).$$

R is one-to-one and the inverse mapping

$$R^{-1}: \mathcal{S}'_{r,\mathrm{cone}}(\mathbb{R}^n \times \mathbb{R}^{n+1}) \to \mathcal{S}'_e(\mathbb{R}^{n+1})$$

is continuous. Moreover, if $g=Rf$ and if $\hat{f}$ or $\hat{g}$ are in some open set equal to ordinary integrable functions $\hat{f}(\sigma, \omega)$ or $\hat{g}(\sigma, \rho)$, then

$$\hat{g}(\sigma, \rho)=\begin{cases}\dfrac{2}{\omega_n}\dfrac{1}{\rho^{n-1}}\dfrac{\hat{f}(\sigma, \sqrt{\rho^2-|\sigma|^2})}{\sqrt{\rho^2-|\sigma|^2}} & \text{for } \rho>|\sigma|,\\ 0 & \text{for } 0 \leqq \rho<|\sigma|\end{cases}$$

and conversely we have the inversion formula

$$\hat{f}(\sigma, \omega)=\frac{\omega_n}{2}\cdot|\omega|(|\sigma|^2+\omega^2)^{(n-1)/2}\hat{g}(\sigma, \sqrt{|\sigma|^2+\omega^2}).$$

The proof of Theorem 2.1 will be divided into several steps arranged in the paragraphs (a) through (i) below.

(a) PROPOSITION 2.2. *There exists a linear and continuous dual mapping*

$$R^*: \mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1}) \to \mathcal{S}_e(\mathbb{R}^{n+1})$$

such that $\langle Rf, \varphi\rangle = \langle f, R^*\varphi\rangle$ *for all* $f \in \mathcal{S}_e(\mathbb{R}^{n+1})$ *and* $\varphi \in \mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$. *Further*

$$R^*\varphi(x, y) = \int_{\mathbb{R}^n} \varphi(z, \sqrt{|z-x|^2+y^2})\, dz$$

and

$$(R^*\varphi)\hat{}(\sigma, \omega) = \hat{\varphi}(\sigma, \sqrt{|\sigma|^2+\omega^2})$$

for all (x, y) *and* $(\sigma, \omega) \in \mathbb{R}^{n+1}$.

Proof.

$$\begin{aligned}
\langle Rf, \varphi\rangle &= \int_{\mathbb{R}^n} \int_0^\infty \bar{\varphi}(x, r) r^n\, dr\, dx \int_{S^n} f(x + r\xi, r\eta)\, dS_n(\xi, \eta) \\
&= \int_{\mathbb{R}^n} dx \int_0^\infty \int_{S^n} \bar{\varphi}(x, r) f(x + r\xi, r\eta) r^n\, dr\, dS_n(\xi, \eta) \\
&= \{z = r\xi,\ y = r\eta\} \\
&= \int_{\mathbb{R}^n} dx \int_{\mathbb{R}^{n+1}} \bar{\varphi}(x, \sqrt{|z|^2+y^2}) f(x+z, y)\, dz\, dy \\
&= \int_{\mathbb{R}^n} dx \int_{\mathbb{R}^{n+1}} \bar{\varphi}(x, \sqrt{|z-x|^2+y^2}) f(z, y)\, dz\, dy \\
&= \int_{\mathbb{R}^{n+1}} f(z, y)\, dz\, dy \int_{\mathbb{R}^n} \bar{\varphi}(x, \sqrt{|x-z|^2+y^2})\, dx \\
&= \int_{\mathbb{R}^{n+1}} f(x, y)\, dx\, dy \int_{\mathbb{R}^n} \bar{\varphi}(z, \sqrt{|z-x|^2+y^2})\, dz = \langle f, R^*\varphi\rangle
\end{aligned}$$

where, by definition, we take

$$R^*\varphi(x, y) = \int_{\mathbb{R}^n} \varphi(z, \sqrt{|z-x|^2+y^2})\, dz.$$

The changing of the order of integration is easily justified by the fact that $f \in \mathcal{S}_e$ and $\varphi \in \mathcal{S}_r$. Obviously $R^*\varphi \in \mathcal{S}'_e(\mathbb{R}^{n+1})$. We have to prove that, actually, $R^*\varphi \in \mathcal{S}_e(\mathbb{R}^{n+1})$ and that the mapping R^* is continuous. To this end we first calculate the Fourier transform $(R^*\varphi)\hat{}(\sigma, \omega)$.

$$\begin{aligned}
(R^*\varphi)\hat{}(\sigma, \omega) &= \int_{\mathbb{R}^{n+1}} e^{-i(\langle \sigma, x\rangle + \omega y)}\, dx\, dy \int_{\mathbb{R}^n} \varphi(z, \sqrt{|x-z|^2+y^2})\, dz \\
&= \int_{\mathbb{R}^n} e^{-i\langle \sigma, z\rangle}\, dz \int_{\mathbb{R}^{n+1}} e^{-i(\langle \sigma, x-z\rangle + \omega y)} \varphi(z, \sqrt{|x-z|^2+y^2})\, dx\, dy \\
&= \int_{\mathbb{R}^n} e^{-i\langle \sigma, z\rangle}\, dz \int_{\mathbb{R}^{n+1}} e^{-i(\langle \sigma, x\rangle + \omega y)} \varphi(z, \sqrt{|x|^2+y^2})\, dx\, dy \\
&= \hat{\varphi}(\sigma, \sqrt{|\sigma|^2+\omega^2}).
\end{aligned}$$

Here, again the change in the order of integration is easily justified since the integrals are absolutely convergent.

LEMMA 2.3. *If* $\hat{\varphi} \in \mathcal{S}_r$ *then* $(R^*\varphi)^\wedge \in \mathcal{S}_e$. *The mapping* $\mathcal{F} \circ R^* \circ \mathcal{F}^{-1}: \mathcal{S}_r \to \mathcal{S}_e$ *is continuous.*

Proof. $\hat{\varphi}(\sigma, \rho)$ may be written $\hat{\varphi}(\sigma, \rho) = \hat{\psi}(\sigma, z, \omega)$ where $\hat{\psi} \in \mathcal{S}(\mathbb{R}^{2n+1})$, $\sigma \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, $\omega \in \mathbb{R}$ and $\rho = \sqrt{|z|^2 + \omega^2}$. Then $(R^*\varphi)^\wedge(\sigma, \omega) = \hat{\varphi}(\sigma, \sqrt{|\sigma|^2 + \omega^2}) = \hat{\psi}(\sigma, \sigma, \omega)$. Using the chain rule, it is easy to verify that $(R^*\varphi)^\wedge \in \mathcal{S}_e$ and that the mapping $\mathcal{F} \circ R^* \circ \mathcal{F}^{-1}$ is continuous.

It follows immediately from the lemma that $R^*\varphi \in \mathcal{S}_e$ whenever $\varphi \in \mathcal{S}_r$ and that the mapping

$$R^*: \mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1}) \to \mathcal{S}_e(\mathbb{R}^{n+1})$$

is continuous. This finishes the proof of Proposition 2.2.

(b) For $f \in \mathcal{S}'_e(\mathbb{R}^{n+1})$ we now define $Rf \in \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$ by the equality

$$\langle Rf, \varphi \rangle = \langle f, R^*\varphi \rangle, \quad \varphi \in \mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1}),$$

i.e., by $Rf = f \circ R^*$. From the continuity of R^* it is clear that Rf is an element of $\mathcal{S}'_r$ for every $f \in \mathcal{S}'_e$. The continuity of the extended mapping

$$R: \mathcal{S}'_e \to \mathcal{S}'_r$$

follows trivially from the definition.

(c) We will also need the following extension lemma of Whitney type.

EXTENSION LEMMA 2.4. *There exists a continuous linear mapping*

$$\mathcal{E}: \mathcal{S}_e(\mathbb{R}^{n+1}) \to \mathcal{S}_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$$

such that for all $\psi \in \mathcal{S}_e(\mathbb{R}^{n+1})$

$$(\mathcal{E}\psi)^\wedge(\sigma, \rho) = \hat{\psi}(\sigma, \sqrt{\rho^2 - |\sigma|^2}) \quad \text{for } \rho \geqq |\sigma|.$$

Postponing the proof of this lemma, we first note that the following corollary follows immediately.

COROLLARY 2.5. $(R^*\mathcal{E}\psi)^\wedge(\sigma, \omega) = \hat{\psi}(\sigma, \omega)$ *for all* $\psi \in \mathcal{S}_e$, *i.e.,* $R^*\mathcal{E} = id_{\mathcal{S}_e}$. *The mapping* $R^*: \mathcal{S}_r \to \mathcal{S}_e$ *is onto. If* $\varphi \in \mathcal{S}_r$ *then* $(\mathcal{E}R^*\varphi)^\wedge(\sigma, \rho) = \hat{\varphi}(\sigma, \rho)$ *for* $\rho \geqq |\sigma|$.

(d) PROPOSITION 2.6. *The mapping*

$$R: \mathcal{S}'_e(\mathbb{R}^{n+1}) \to \mathcal{S}'_r(\mathbb{R}^n \times \mathbb{R}^{n+1})$$

is one-to-one.

Proof. Suppose that $Rf = 0$ and let $\psi \in \mathcal{S}_e$ be arbitrary. Then by Corollary 2.5

$$\langle f, \psi \rangle = \langle f, R^*\mathcal{E}\psi \rangle = \langle Rf, \mathcal{E}\psi \rangle = 0.$$

Consequently $f = 0$ which finishes the proof.

(e) PROPOSITION 2.7. *If* $f \in \mathcal{S}'_e$ *and* $g = Rf$ *then* $\langle f, \psi \rangle = \langle g, \mathcal{E}\psi \rangle$ *for all* $\psi \in \mathcal{S}_e$.

This was already proved in the previous proposition.

(f) PROPOSITION 2.8. *For every* $f \in \mathcal{S}'_e$ *we have* $g = Rf \in \mathcal{S}'_{r,\text{cone}}$.

Proof. Let φ be arbitrary such that $\varphi \in \mathcal{S}_r$ and supp $\hat{\varphi} \subset \{(\sigma, \rho): 0 \leqq \rho < |\sigma|\}$. Then $\langle Rf, \varphi \rangle = \langle f, R^*\varphi \rangle = (2\pi)^{-(n+1)} \langle \hat{f}, (R^*\varphi)^\wedge \rangle$. Since $(R^*\varphi)^\wedge(\sigma, \omega) = \hat{\varphi}(\sigma, \sqrt{|\sigma|^2 + \omega^2}) = 0$ for all $(\sigma, \omega) \in \mathbb{R}^{n+1}$ we deduce that $\langle Rf, \varphi \rangle = 0$ which finishes the proof.

LEMMA 2.9. *If* $\hat{g} \in \mathcal{S}'_r$, $\hat{h} \in \mathcal{S}_r$ *and if* supp $\hat{g} \subset \{(\sigma, \rho): \rho \geqq |\sigma|\}$ *and* supp $\hat{h} \subset \{(\sigma, \rho): \rho \leqq |\sigma|\}$ *then* $\langle \hat{g}, \hat{h} \rangle = 0$.

Proof. Choose $\varphi \in C_0^\infty(\mathbb{R}^{n+1})$ such that $\varphi(0)=1$ and take $\varphi_n(x)=\varphi(x/n)$. Then, if we let $\hat{h}_n=\varphi_n\hat{h}$, $\hat{h}_n\to\hat{h}$ in $\mathcal{S}(\mathbb{R}^{n+1})$ and supp $\hat{h}_n$ is compact. By [8, Thm. 2.3.3], $\langle\hat{g},\hat{h}_n\rangle=0$. Consequently $\langle\hat{g},\hat{h}\rangle=\lim_{n\to\infty}\langle\hat{g},\hat{h}_n\rangle=0$ and the proof is complete.

(g) PROPOSITION 2.10. *For every $g\in\mathcal{S}'_{r,\text{cone}}$ there exists an $f\in\mathcal{S}'_e$ such that $g=Rf$.*

Proof. Let $g\in\mathcal{S}'_{r,\text{cone}}$ be given. Now, since the mapping $\mathcal{E}$ is continuous, the relation

$$\langle f,\psi\rangle=\langle g,\mathcal{E}\psi\rangle$$

with $\psi\in\mathcal{S}_e$ arbitrary, defines a distribution $f\in\mathcal{S}'_e$. We shall prove that $g=Rf$. Using that supp $\hat{g}\subset\{(\sigma,\rho)\colon \rho\geqq|\sigma|\}$, Corollary 2.5 and Lemma 2.9 we have

$$\langle g,\varphi\rangle=(2\pi)^{-(n+1)}\langle\hat{g},\hat{\varphi}\rangle=(2\pi)^{-(n+1)}\langle\hat{g},(\mathcal{E}R^*\varphi)\hat{}\,\rangle=\langle g,\mathcal{E}R^*\varphi\rangle.$$

By the definition of f this last expression equals $\langle f,R^*\varphi\rangle=\langle Rf,\varphi\rangle$. Therefore $\langle g,\varphi\rangle=\langle Rf,\varphi\rangle$ for all $\varphi\in\mathcal{S}_r$, i.e., $g=Rf$.

PROPOSITION 2.11. *If $g_n\in\mathcal{S}'_{r,\text{cone}}$, $g_n\to g\in\mathcal{S}'_r$ and if $Rf_n=g_n$ then $f_n\to f$ where $Rf=g$, i.e., the inverse mapping is continuous.*

Proof. By Proposition 2.7 $\langle f_n,\psi\rangle=\langle g_n,\mathcal{E}\psi\rangle$ for all $\psi\in\mathcal{S}_r$. Since $\mathcal{S}'_{r,\text{cone}}$ is closed $g\in\mathcal{S}'_{r,\text{cone}}$ and $g=Rf$ for some $f\in\mathcal{S}'_e$. Now

$$\langle f_n,\psi\rangle\to\langle g,\mathcal{E}\psi\rangle=\langle f,\psi\rangle.$$

(h) Let $U\subset\{(\sigma,\rho)\colon\rho>|\sigma|\}\subset\mathbb{R}^{2n+1}$ be an open set of the form $U=\{(\sigma,\rho)\colon\sigma\in V,0<a<\rho<b\}$ where $V\subset\mathbb{R}^n$ is open. Let $U'\subset\mathbb{R}^{n+1}$ be the image of U under the mapping $(\sigma,\rho)\mapsto(\sigma,\sqrt{\rho^2-|\sigma|^2})$. It is easy to see that U' is open and that U is the inverse image of U' under the same mapping. Further consider the mappings $\hat{\varphi}\mapsto(R^*\varphi)\hat{}\,=\hat{\psi}$ and $\hat{\psi}\mapsto(\mathcal{E}\psi)\hat{}\,=\hat{\varphi}$. It is rather easy to see that they establish a one-to-one correspondence between all $\hat{\varphi}\in\mathcal{S}_r$ with supp $\hat{\varphi}\subset U$ and all $\hat{\psi}\in\mathcal{S}_e$ with supp $\hat{\psi}\subset U'$.

Now suppose that the distribution $\hat{g}=(Rf)\hat{}$ is on U equal to some integrable function $\hat{g}(\sigma,\rho)$. Then

$$\begin{aligned}\langle\hat{g},\hat{\varphi}\rangle&=\omega_n\int_{\mathbb{R}^n}\int_{|\sigma|}^{\infty}\hat{g}(\sigma,\rho)\bar{\hat{\varphi}}(\sigma,\rho)\rho^n\,d\rho\,d\sigma=\{\text{substitute }\omega=\sqrt{\rho^2-|\sigma|^2}\}\\&=\frac{\omega_n}{2}\int_{\mathbb{R}^{n+1}}\hat{g}(\sigma,\sqrt{|\sigma|^2+\omega^2})(|\sigma|^2+\omega^2)^{(n-1)/2}|\omega|\bar{\hat{\varphi}}(\sigma,\sqrt{|\sigma|^2+\omega^2})\,d\sigma\,d\omega\\&=\langle\hat{f},(R^*\varphi)\hat{}\,\rangle.\end{aligned}$$

Taking $\psi=R^*\varphi$ we have

$$\langle\hat{f},\hat{\psi}\rangle=\frac{\omega_n}{2}\int_{\mathbb{R}^{n+1}}\hat{g}(\sigma,\sqrt{|\sigma|^2+\omega^2})(|\sigma|^2+\omega^2)^{(n-1)/2}|\omega|\bar{\hat{\psi}}(\sigma,\omega)\,d\sigma\,d\omega$$

where $\hat{\psi}$ with supp $\hat{\psi}\subset U'$ is arbitrary. We conclude that

$$\hat{f}(\sigma,\omega)=\frac{\omega_n}{2}|\omega|(|\sigma|^2+\omega^2)^{(n-1)/2}\hat{g}(\sigma,\sqrt{|\sigma|^2+\omega^2})$$

for $(\sigma,\omega)\in U'$. This is our inversion formula. Next suppose that the distribution $\hat{f}$ is on U' equal to some integrable function $\hat{f}(\sigma,\omega)$. By Proposition 2.7 we have, if supp $\hat{\psi}\subset U'$,

$$\begin{aligned}\langle\hat{g},(\mathcal{E}\psi)\hat{}\,\rangle=\langle\hat{f},\hat{\psi}\rangle&=\int_{\mathbb{R}^{n+1}}\hat{f}(\sigma,\omega)\bar{\hat{\psi}}(\sigma,\omega)\,d\sigma\,d\omega=\{\text{substitute }\rho=\sqrt{|\sigma|^2+\omega^2}\}\\&=\frac{2}{\omega_n}\cdot\omega_n\int_{\mathbb{R}^n}\int_{|\sigma|}^{\infty}\frac{\hat{f}(\sigma,\sqrt{\rho^2-|\sigma|^2})}{\sqrt{\rho^2-|\sigma|^2}}\,\frac{\bar{\hat{\psi}}(\sigma,\sqrt{\rho^2-|\sigma|^2})}{\rho^{n-1}}\rho^n\,d\rho\,d\sigma.\end{aligned}$$

Since $\hat{\psi}(\sigma, \sqrt{\rho^2-|\sigma|^2}) = (\mathscr{E}\psi)\hat{}(\sigma, \rho)$ we have, introducing $\varphi = \mathscr{E}\psi$,

$$\langle \hat{g}, \hat{\varphi} \rangle = \frac{2}{\omega_n} \omega_n \int_{\mathbb{R}^n} \int_{|\sigma|}^{\infty} \frac{\hat{f}(\sigma, \sqrt{\rho^2-|\sigma|^2})}{\rho^{n-1}\sqrt{\rho^2-|\sigma|^2}} \bar{\varphi}(\sigma, \rho) \rho^n \, d\rho \, d\sigma$$

whence we conclude that

$$\hat{g}(\sigma, \rho) = \frac{2}{\omega_n} \frac{\hat{f}(\sigma, \sqrt{\rho^2-|\sigma|^2})}{\rho^{n-1}\sqrt{\rho^2-|\sigma|^2}} \quad \text{for } (\sigma, \rho) \in U.$$

(i) To complete the proof of Theorem 2.1 it remains to prove the extension lemma. For that purpose we need a few sublemmata.

LEMMA 2.12. *The mapping* $A: \mathscr{S}_e(\mathbb{R}^{n+1}) \to C^\infty(\mathbb{R}^{n+1})$, *given by* $A\varphi(x, y) = (1/y)\varphi'_y(x, y)$ *has its range in* $\mathscr{S}_e(\mathbb{R}^{n+1})$ *and is a continuous mapping,*

$$A: \mathscr{S}_e(\mathbb{R}^{n+1}) \to \mathscr{S}_e(\mathbb{R}^{n+1}).$$

Proof. Considering Fourier transforms we may equivalently prove that the mapping

$$\hat{\varphi}(\sigma, \omega) \to \int_{\omega}^{\infty} s\hat{\varphi}(\sigma, s) \, ds$$

is a continuous mapping from $\mathscr{S}_e$ *to* $\mathscr{S}_e$. This is almost obvious.

Let us now consider functions $\psi(\sigma, t)$ defined for $t > |\sigma|^2$ in $\mathbb{R}^{n+1}$, which are C^∞ and such that the norms

$$\|\psi\|_{\mathrm{par},M} = \sup \left\{ (t^2+|\sigma|^2)^m \left| \frac{\partial^{\alpha+|\beta|}}{\partial t^\alpha \, \partial \sigma^\beta} \psi(\sigma, t) \right| : t > |\sigma|^2, 0 \leqq m \leqq M, 0 \leqq \alpha + |\beta| \leqq M \right\}$$

are finite. The linear space of such functions with the topology given by these norms is denoted by $\mathscr{S}_{\mathrm{par}}(\mathbb{R}^{n+1})$.

LEMMA 2.13. *The mapping*

$$B: \mathscr{S}_e(\mathbb{R}^{n+1}) \to \mathscr{S}_{\mathrm{par}}(\mathbb{R}^{n+1})$$

given by $B\varphi(\sigma, t) = \varphi(\sigma, \sqrt{t-|\sigma|^2})$ *is linear and continuous.*

Proof. If $\psi(\sigma, t) = B\varphi(\sigma, t) = \varphi(\sigma, \sqrt{t-|\sigma|^2})$ then

$$\frac{\partial \psi}{\partial t}(\sigma, t) = \frac{1}{2\sqrt{t-|\sigma|^2}} \frac{\partial \varphi}{\partial \omega}(\sigma, \sqrt{t-|\sigma|^2}) = \frac{1}{2}\left(\frac{1}{\omega} \frac{\partial \varphi}{\partial \omega}(\sigma, \omega) \right)_{\omega = \sqrt{t-|\sigma|^2}}.$$

Consequently $(\partial/\partial t)(B\varphi) = \frac{1}{2} BA\varphi$ where A is the mapping in Lemma 2.12. Further

$$\frac{\partial \psi}{\partial \sigma_i} = \frac{\partial \varphi}{\partial \sigma_i}(\sigma, \sqrt{t-|\sigma|^2}) - \frac{\sigma_i}{\sqrt{t-|\sigma|^2}} \frac{\partial \varphi}{\partial \omega}(\sigma, \sqrt{t-|\sigma|^2})$$

i.e. $(\partial/\partial\sigma_i)(B\varphi) = B((\partial\varphi/\partial\sigma_i) - \sigma_i A\varphi) = BC_i\varphi$ where the mapping $C_i: \mathscr{S}(\mathbb{R}^{n+1}) \to \mathscr{S}(\mathbb{R}^{n+1})$ defined by $C_i\varphi(\sigma, \omega) = \partial\varphi/\partial\sigma_i(\sigma, \omega) - \sigma_i A\varphi(\sigma, \omega)$ is continuous. Further we see that $AC_i = C_i A$ which implies that

$$\frac{\partial^{\alpha+|\beta|}}{\partial t^\alpha \partial \sigma^\beta} B\varphi = \frac{1}{2^\alpha} BA^\alpha C^\beta \varphi.$$

Here $C^\beta = C_1^{\beta_1} C_2^{\beta_2} \cdots C_n^{\beta_n}$. Then

$$(t^2+|\sigma|^2)^m \frac{\partial^{\alpha+|\beta|}}{\partial t^\alpha \partial \sigma^\beta} \psi(\sigma, t) = \frac{1}{2^\alpha} B(((\omega^2+|\sigma|^2)^2+|\sigma|^2)^m A^\alpha C^\beta)\varphi.$$

Using the estimate

$$\sup_{t>|\sigma|^2} |B\varphi(\sigma, t)| < K_0 \sup_{(\sigma,\omega)\in\mathbb{R}^{n+1}} |\varphi(\sigma, \omega)|$$

and the continuity of A and C we obtain

$$\sup\left\{(t^2+|\sigma|^2)^m \left|\frac{\partial^{\alpha+|\beta|}}{\partial t^\alpha \partial\sigma^\beta}\psi(\sigma, t)\right| : t>|\sigma|^2, 0\leqq\alpha+|\beta|\leqq M, 0\leqq m\leqq M\right\}$$
$$\leqq K_M \sup\left\{(|\sigma|^2+\omega^2)^m \left|\frac{\partial^{\alpha+|\beta|}}{\partial\omega^\alpha\partial\sigma^\beta}\varphi(\sigma,\omega)\right| : (\sigma,\omega)\in\mathbb{R}^{n+1},\right.$$
$$\left. 0\leqq\alpha+|\beta|\leqq M', 0\leqq m\leqq M'\right\}$$

for some constants K_M and M'. This proves that B is continuous.

LEMMA 2.14. *There exists a linear continuous mapping*

$$\mathcal{G}: \mathcal{S}_{\text{par}}(\mathbb{R}^{n+1}) \to \mathcal{S}(\mathbb{R}^{n+1})$$

such that $\mathcal{G}\psi(\sigma, t) = \psi(\sigma, t)$ *for* $t>|\sigma|^2$.

Proof. Following the method used in Chapter VI in the book by Stein [12] we first introduce a function $h(\lambda)$ with the following properties. $h(\lambda)$ is defined for $\lambda\geqq 1$, $h(\lambda)$ is rapidly decreasing at ∞, i.e., $h(\lambda)=O(\lambda^{-N})$ as $\lambda\to\infty$ for all N and, finally, $\int_1^\infty h(\lambda)\,d\lambda=1$, $\int_1^\infty \lambda^n h(\lambda)\,d\lambda=0$ for $n=1, 2, \cdots$. Then define the extension $\mathcal{G}\psi$ by

$$\mathcal{G}\psi(\sigma, t)=\begin{cases}\displaystyle\int_1^\infty \psi(\sigma, t+2\lambda(|\sigma|^2-t))h(\lambda)\,d\lambda & \text{for } t<|\sigma|^2,\\ \psi(\sigma, t) & \text{for } t>|\sigma|^2,\\ \displaystyle\lim_{t\searrow|\sigma|^2}\psi(\sigma, t) & \text{for } t=|\sigma|^2.\end{cases}$$

Now, by the arguments in [12, pp. 185–186] it follows that $\psi\in\mathcal{S}_{\text{par}}(\mathbb{R}^{n+1})\Rightarrow\mathcal{G}\psi\in C^\infty(\mathbb{R}^{n+1})$. To prove the continuity we proceed as follows.

If $t\geqq|\sigma|^2$ then $\psi(\sigma, t)=\psi(\sigma, t+)$.

If $t<|\sigma|^2$ and $|\alpha|+\beta=M$ then it easily follows that

$$\frac{\partial^{|\alpha|+\beta}}{\partial\sigma^\alpha\partial t^\beta}\mathcal{G}\psi(\sigma, t)=\sum_{m=0}^{M}\int_1^\infty p_m\left(\sigma, \frac{\partial}{\partial\sigma}, \frac{\partial}{\partial t}\right)\psi(\sigma, t^*)\lambda^m h(\lambda)\,d\lambda$$

where $t^*=t^*(\sigma, t, \lambda)=t+2\lambda(|\sigma|^2-t)$ and $p_m(\cdot,\cdot,\cdot)$ is a three-variable polynomial of degree less than or equal to m. Multiplying by $(|\sigma|^2+t^2)^{N/2}$ and taking the supremum when $t<|\sigma|^2$ we obtain, for some constant C_M

$$\|\mathcal{G}\psi\|_N \leqq C_M \|\psi\|_{\text{par},M+N} \sum_{m=0}^{M}\int_1^\infty \lambda^m|h(\lambda)|\,d\lambda.$$

Since h is rapidly decreasing, all the integrals are finite. This proves the continuity of the mapping $\mathcal{G}: \mathcal{S}_{\text{par}}(\mathbb{R}^{n+1})\to\mathcal{S}(\mathbb{R}^{n+1})$.

LEMMA 2.15. *The mapping*

$$K: \mathcal{S}(\mathbb{R}^{n+1})\to\mathcal{S}_r(\mathbb{R}^n\times\mathbb{R}^{n+1})$$

given by $K\psi(\sigma, \rho)=\psi(\sigma, \rho^2)$ *is linear and continuous.*

Proof. If $\rho^2=|\xi|^2$, $\xi\in\mathbb{R}^{n+1}$ and $\varphi(\sigma, \xi)=\psi(\sigma, \rho^2)$ then

$$\frac{\partial^{|\alpha|+|\beta|}}{\partial\sigma^\alpha\partial\xi^\beta}\varphi(\sigma, \xi)=\sum_{0\leqq l\leqq|\beta|}P_{\alpha,l}(\xi)\frac{\partial^{|\alpha|+l}}{\partial\sigma^\alpha\partial\omega^l}\psi(\sigma, |\xi|^2)$$

where $P_{\alpha,l}(\xi)$ are polynomials in $\xi_1, \xi_2, \cdots, \xi_{n+1}$. From this it follows that K is continuous.

Now combining Lemmata 2.13, 2.14 and 2.15, we see that the mapping

$$\mathscr{E} = \mathscr{F} \circ K \circ \mathscr{G} \circ B \circ \mathscr{F}^{-1}$$

satisfies the requirements of the extension lemma. This finally completes the proof of Theorem 2.1.

Remark 2.16. In case $\hat{f}$ or $\hat{g}$ are not locally equal to integrable functions, an inversion formula is still supplied by the relation

$$\langle f, \varphi \rangle = \langle g, \mathscr{E}\varphi \rangle$$

given in Proposition 2.7. This may also be written $\langle f, \varphi \rangle = \langle \mathscr{E}^* g, \varphi \rangle$ or

$$f = \mathscr{E}^* g \quad \text{where } \mathscr{E}^* : \mathscr{S}'_r \to \mathscr{S}'_e.$$

Consequently $R\mathscr{E}^* g = g \Leftrightarrow \hat{g} \in \mathscr{S}'_{r,\text{cone}}$ and $\mathscr{E}^* Rf = f$ for all $f \in \mathscr{S}'_e$.

3. A Sobolev estimate.

DEFINITION. $\mathscr{H}^\alpha(\mathbb{R}^n)$ is the set of all distributions f in $\mathscr{S}'(\mathbb{R}^n)$ such that $\hat{f}$ has a representative which is a locally integrable function with the norm $\|f\|_\alpha = \{\int_{\mathbb{R}^n} (1+|\xi|^2)^\alpha |\hat{f}(\xi)|^2 \, d\xi\}^{1/2} < \infty$.

THEOREM 3.1. *If $g = Rf$ then*

$$\|f\|_\alpha < \sqrt{\frac{\omega_n}{2}} \|g\|_{\alpha+1/2}.$$

Proof. By the inversion formula

$$\hat{f}(\sigma, \omega) = \frac{\omega_n}{2} |\omega| (|\sigma|^2 + \omega^2)^{(n-1)/2} \hat{g}(\sigma, \sqrt{|\sigma|^2 + \omega^2})$$

we have

$$\begin{aligned}
&\int_{\mathbb{R}^{n+1}} (1+|\sigma|^2+\omega^2)^\alpha |\hat{f}(\sigma,\omega)|^2 \, d\sigma \, d\omega \\
&= \frac{\omega_n^2}{4} \int_{\mathbb{R}^{n+1}} (1+|\sigma|^2+\omega^2)^\alpha \omega^2 (|\sigma|^2+\omega^2)^{n-1} |\hat{g}(\sigma, \sqrt{|\sigma|^2+\omega^2})|^2 \, d\sigma \, d\omega \\
&= \{\text{substitute } \rho = \sqrt{|\sigma|^2+\omega^2}\} \\
&= \frac{\omega_n^2}{2} \int_{\mathbb{R}^n} \int_{|\sigma|}^\infty (1+\rho^2)^\alpha (\rho^2 - |\sigma|^2) \rho^{n-1} \frac{|\hat{g}(\sigma,\rho)|^2}{\sqrt{\rho^2-|\sigma|^2}} \rho \, d\rho \, d\sigma \\
&= \frac{\omega_n^2}{2} \int_{\mathbb{R}^n} \int_{|\sigma|}^\infty (1+\rho^2)^\alpha \sqrt{\rho^2-|\sigma|^2} |\hat{g}(\sigma,\omega)|^2 \rho^n \, d\rho \, d\sigma \\
&\leqq \frac{\omega_n^2}{2} \int_{\mathbb{R}^n} \int_{|\sigma|}^\infty (1+\rho^2+|\sigma|^2)^{\alpha+1/2} |\hat{g}(\sigma,\rho)|^2 \rho^n \, d\rho \, d\sigma \\
&= \frac{\omega_n}{2} \|g\|^2_{\alpha+1/2}.
\end{aligned}$$

Before proceeding to the local problem we remark that several reformulations of the inversion formula

$$\hat{f}(\sigma, \omega) = \frac{\omega_n}{2} |\omega| (|\sigma|^2 + \omega^2)^{(n-1)/2} \hat{g}(\sigma, \sqrt{|\sigma|^2 + \omega^2})$$

are possible. Taking inverse Fourier transforms and using the formula of Proposition 2.2, we obtain $f = c_n H_y(\partial/\partial y)\Delta^{(n-1)/2} R^* g$ which is also given by Fawcett [3]. Here H_y is the Hilbert transform with respect to y, Δ is the Laplacian $\Delta = \Delta_x + (\partial^2/\partial y^2)$ and c_n is a constant. However, if we introduce the pseudodifferential operator K defined by $(Kg)\hat{}(\sigma, \rho) = \sqrt{\rho^2 - |\sigma|^2}\, \rho^{n-1} \hat{g}(\sigma, \rho)$ then also $f = c_n R^* K g$.

Here, as is easily seen, K is a continuous mapping from the subspace $\mathcal{H}^\alpha \cap \mathcal{S}'_{r,\text{cone}}$ to $\mathcal{H}^{\alpha+n} \cap \mathcal{S}'_r$ and R^* is a continuous mapping from $\mathcal{H}^{\alpha+n} \cap \mathcal{S}'_r$ to $\mathcal{H}^{\alpha+1/2}$.

This formula might be more useful for numerical purposes than Fawcett's. One difficulty in using any of these formulas numerically will probably be the fact that the calculation of R^* requires an integration over an unbounded domain.

4. Uniqueness for the local problem. In this section we will extend a theorem of uniqueness for the local problem, proved by Courant and Hilbert [2] for continuous functions f.

Let $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$ and $B > 0$ be given.

THEOREM 4.1. *Let $f \in \mathcal{S}'_e(\mathbb{R}^{n+1})$ and suppose that $g = Rf$ is equal to zero on the open set $U_{B,\varepsilon} = \{(x, r): |x - x_0| < \varepsilon,\ 0 \leqq r < B\} \subset \mathbb{R}^n \times \mathbb{R}^{n+1}$. Then $f = 0$ on the open set $V_B = \{(x, y): |x - x_0|^2 + y^2 < B^2\} \subset \mathbb{R}^{n+1}$. Also $g = 0$ in the open double cone $W_B = \{(x, r): |x - x_0| + r < B\}$.*

We first note that the assumption that f is a *tempered* distribution is inessential in the sense that any distribution defined on an open set $U_{B',\varepsilon'} \supset U_{B,\varepsilon}$ is, when restricted to $U_{B,\varepsilon}$, equal to the restriction of some *tempered* distribution to $U_{B,\varepsilon}$. We also see that we may take $x_0 = 0$ without loss of generality.

LEMMA 4.2. *Let $\varphi \in \mathcal{S}_r$, $\psi = R^*\varphi$, $f \in \mathcal{S}'_e$ and $g = Rf$. Then $g * \varphi = R(f * \psi)$.*

Proof. Consider the continuous mapping $f \mapsto R(f * (R^*\varphi)) - (Rf) * \varphi = h$ from $\mathcal{S}'_e$ to $\mathcal{S}'_r$.

Using the inversion formula of Theorem 2.1 and the formula for $(R^* \varphi)\hat{}$ in Proposition 2.2 it follows that $\hat{h} = 0$ if $f \in \mathcal{S}_r$, i.e., that the mapping is identically zero on the dense subspace $\mathcal{S}_e \subset \mathcal{S}'_e$. By continuity it follows that $R(f * (R^*\varphi)) - (Rf) * \varphi = 0$ for all $f \in \mathcal{S}'_e$ which proves the lemma.

Proof of Theorem 4.1. The first part of the proof is merely a reproduction of Courant's proof and is included for the convenience of the reader. Let $f \in C^\infty(\mathbb{R}^{n+1})$ and $g = Rf$. Obviously $g \in C^\infty(\mathbb{R}^n \times \mathbb{R}^{n+1})$. Take

$$G(x, r) = \omega_n \int_0^r g(x, s) s^n \, ds = \int_{|\xi|^2 + \eta^2 \leqq r^2} f(x + \xi, \eta) \, d\xi \, d\eta,$$

$$\begin{aligned} G'_{x_i}(x, r) &= \int_{|\xi|^2 + \eta^2 \leqq r^2} f'_{x_i}(x + \xi, \eta) \, d\xi \, d\eta \\ &= \int_{|\xi|^2 + \eta^2 \leqq r^2} f'_{\xi_i}(x + \xi, \eta) \, d\xi \, d\eta \\ &= \frac{1}{r} \int_{|\xi|^2 + \eta^2 = r^2} f(x + \xi, \eta) \xi_i \, dS_n. \end{aligned}$$

Now

$$\begin{aligned} R(x_i f)(x, r) &= \frac{1}{r^n \omega_n} \int_{|\xi|^2 + \eta^2 = r^2} f(x + \xi, \eta)(x_i + \xi_i) \, dS_n \\ &= x_i g(x, r) + \frac{r^{1-n}}{\omega_n} G'_{x_i}(x, r) \\ &= x_i g(x, r) + r^{1-n} \frac{\partial}{\partial x_i} \int_0^r g(x, s) s^n \, ds = D_i g(x, r). \end{aligned}$$

Consequently $R(x_i f) = D_i g$ where the linear operator D_i is defined by the previous expression. Repeating, we obtain $R(p(x)f) = p(D)g$ where p is any n-variable polynomial. Further

$$R(p(x)f)(x, r) = \frac{1}{r^n \omega_n} \int_{|\xi|^2+\eta^2=r^2} p(x+\xi) f(x+\xi, \eta)\, dS_n$$

$$= \frac{1}{r^{n-1}\omega_n} \int_{|\xi|\leqq r} p(x+\xi) \frac{f(x+\xi, \sqrt{r^2-|\xi|^2})}{\sqrt{r^2-|\xi|^2}}\, d\xi.$$

Now if $g = 0$ in $U_{B,\varepsilon}$ then obviously $p(D)g = 0$ in $U_{B,\varepsilon}$. Then

$$\int_{|\xi|\leqq r} p(x+\xi) \frac{f(x+\xi, \sqrt{r^2-|\xi|^2})}{\sqrt{r^2-|\xi|^2}}\, d\xi = 0$$

for every fixed point $(x, r) \in U_{B,\varepsilon}$ and every polynomial p. For x and r fixed, select a sequence p_n of polynomials such that $p_n(x+\xi) \to f(x+\xi, \sqrt{r^2-|\xi|^2})$ uniformly for $|\xi| \leqq r$. It follows that $f = 0$ in V_B and that $g = 0$ in W_B.

To proceed to the case when $f \in \mathscr{S}'_e$ we select a sequence of mollifiers $\varphi_n \in \mathscr{S}_r$ such that supp $\varphi_n \subset \{(x, r)\colon |x| < 1/n,\ 0 \leqq r < 1/n\}$ and such that $\varphi_n * g \to g$ for all $g \in \mathscr{S}'_r$. Taking $\psi_n = R^* \varphi_n$ it follows easily that supp $\psi_n \subset \{(x, y)\colon |x| < 2/n,\ 0 \leqq y < 1/n\}$. From Lemma 4.2 we conclude that $R(f * \psi_n) = g * \varphi_n \to g$ in the topology of $\mathscr{S}'_r$. Since R^{-1} is continuous we have $f * \psi_n \to f$ in the topology of $\mathscr{S}'_e$. Moreover, $g * \varphi_n$ and $f * \psi_n$ are in C^∞, $g * \varphi_n = 0$ in a sequence $U_{B_n,\varepsilon_n} \subset U_{B,\varepsilon}$ of open sets such that $B_n \to B$, $\varepsilon_n \to \varepsilon$. Then by the previous result $f * \psi_n = 0$ in the sequence $V_{B_n} \subset V_B$ of open sets. Since $f * \psi_n \to f$ as $n \to \infty$ in the sense of distributions it follows that $f = 0$ in V_B. Then also $g = 0$ in W_B and the proof is complete.

COROLLARY 4.3. *If $g = 0$ in a strip $\{(x, r)\colon |x - x_0| < \varepsilon,\ 0 \leqq r < \infty\}$ then $g = 0$ everywhere.*

Proof. Let $B \to \infty$ in Theorem 4.1.

5. Inversion formulas for the local problem. We will start by investigating the mapping R when f and g are of the form $f(x, y) = F(y)\, e^{i\langle\sigma,x\rangle}$ and $g(x, r) = G(r)\, e^{i\langle\sigma,x\rangle}$. The resulting equations have somewhat different form for even and odd values of n.

THEOREM 5.1. *Let $f(x, y) = F(y)\, e^{i\langle\sigma,x\rangle}$ where $F(y)$ is continuous for $0 \leqq y < B \leqq \infty$. Then $g = Rf$ is of the form $g(x, r) = G(r)\, e^{i\langle\sigma,x\rangle}$ where*

$$(1)\quad \left(\frac{|\sigma|}{2}\right)^{n-1} \frac{\sqrt{\pi}}{2\Gamma((n+1)/2)} r^{n-1} G(r) = \int_0^r F(y)(r^2-y^2)^{(n-2)/4} J_{(n-2)/2}(|\sigma|\sqrt{r^2-y^2})\, dy$$

for $0 \leqq r < A$.

Further $r^k G^{(k)}(r)$ is continuous on $[0, B)$ for $0 \leqq k \leqq n/2$. We have also for $n = 2m$, $m \geqq 1$

$$(2)\quad \frac{d}{dr}\left(\frac{1}{2r}\frac{d}{dr}\right)^{m-1} (r^{2m-1} G(r)) = \frac{2\Gamma(m+\frac{1}{2})}{\sqrt{\pi}} \left\{ F(r) - |\sigma| r \int_0^r F(y) \frac{J_1(|\sigma|\sqrt{r^2-y^2})}{\sqrt{r^2-y^2}}\, dy \right\}$$

and for $n = 2m+1$, $m \geqq 0$

$$(3)\quad \left(\frac{1}{2r}\frac{d}{dr}\right)^m (r^{2m} G(r)) = \frac{2\Gamma(m+1)}{\pi} \int_0^r \frac{\cos(|\sigma|\sqrt{r^2-y^2})}{\sqrt{r^2-y^2}} F(y)\, dy.$$

In particular for $n=1$ we have

$$\frac{\pi}{2}G(r)=\int_0^r F(y)\frac{\cos(\sigma\sqrt{r^2-y^2})}{\sqrt{r^2-y^2}}\,dy \tag{4}$$

and for $n=2$

$$rG(r)=\int_0^r F(y)J_0(|\sigma|\sqrt{r^2-y^2})\,dy_1, \tag{5}$$

$$\frac{d}{dr}(rG(r))=F(r)-r|\sigma|\int_0^r F(y)\frac{J_1(|\sigma|\sqrt{r^2-y^2})}{\sqrt{r^2-y^2}}\,dy. \tag{6}$$

Proof.

$$\begin{aligned}
g(x,r)&=\frac{1}{\omega_n r^n}\int_{|\xi|^2+\eta^2=r^2} F(\eta)\,e^{i\langle\sigma,x+\xi\rangle}\,dS_n(\xi,\eta)\\
&=\frac{1}{\omega_n r^n}\,e^{i\langle\sigma,x\rangle}\int_{|\xi|^2+\eta^2=r^2} F(\eta)\,e^{i\langle\sigma,\xi\rangle}\,dS_n(\xi,\eta)\\
&=\left\{dS_n=\frac{r}{|\eta|}\,d\xi,\ |\eta|=\sqrt{r^2-|\xi|^2}\right\}\\
&=\frac{2}{\omega_n r^{n-1}}\,e^{i\langle\sigma,x\rangle}\int_{|\xi|\leqq r}\frac{F(\sqrt{r^2-|\xi|^2})}{\sqrt{r^2-|\xi|^2}}\,e^{i\langle\sigma,\xi\rangle}\,d\xi.
\end{aligned}$$

Hence

$$\frac{\omega_n}{2}r^{n-1}G(r)=\int_{|\xi|\leqq r}\frac{F(\sqrt{r^2-|\xi|^2})}{\sqrt{r^2-|\xi|^2}}\,e^{i\langle\sigma,\xi\rangle}\,d\xi.$$

For $n=1$ we easily obtain (4). For $n\geqq 2$ we introduce polar coordinates $\xi=sx$, $x\in S^{n-1}$, $0\leqq s\leqq r$, $d\xi=s^{n-1}\,ds\,dS_{n-1}(x)$. Then

$$\frac{\omega_n}{2}r^{n-1}G(r)=\int_0^r\frac{F(\sqrt{r^2-s^2})}{\sqrt{r^2-s^2}}s^{n-1}\,ds\int_{S^{n-1}}e^{is\langle\sigma,x\rangle}\,dS_{n-1}(x).$$

Taking $\langle\sigma,x\rangle=|\sigma|\cos v$ we have by Fubini's theorem

$$\begin{aligned}
\int_{S^{n-1}}e^{is\langle\sigma,x\rangle}\,dS_{n-1}(x)&=\int_0^\pi e^{is|\sigma|\cos v}\omega_{n-2}(\sin v)^{n-2}\,dv\\
&=2\omega_{n-2}\int_0^{\pi/2}\cos(s|\sigma|\cos v)(\sin v)^{n-2}\,dv\\
&=\text{(see Magnus and Oberhettinger [10])}\\
&=\omega_{n-2}\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)J_{(n-2)/2}(s|\sigma|)\left(\frac{2}{|\sigma|}\right)^{n/2-1}.
\end{aligned}$$

If we insert this in the previous formula and use that $\omega_n/\omega_{n-2}=2\pi(n-1)$ we obtain (1).

Next if we consider $H(r)=r^{(n-1)/2}G(\sqrt{r})$ and $K(y)=F(\sqrt{y})/2\sqrt{y}$ (1) may be rewritten

$$\left(\frac{|\sigma|}{2}\right)^{(n-2)/2}\frac{\sqrt{\pi}}{2\Gamma((n+1)/2)}H(r)=\int_0^r K(s)(r-s)^{(n-2)/4}J_{(n-2)/4}(|\sigma|\sqrt{r-s})\,ds.$$

Taking Laplace transforms we have (see [10])

$$\frac{\sqrt{\pi}}{2\Gamma((n+1)/2)}\tilde{H}(p)=\tilde{K}(p)\frac{e^{-|\sigma|^2/4p}}{p^{n/2}}.$$

If $n=2m$ then

$$\frac{\sqrt{\pi}}{2\Gamma(m+\frac{1}{2})}p^m\tilde{H}(p)=\tilde{K}(p)\,e^{-|\sigma|^2/4p}.$$

Now

$$\frac{e^{-|\sigma|^2/4p}}{p}=\mathscr{L}(J_0(|\sigma|\sqrt{t}))$$

and so

$$e^{-|\sigma|^2/4p}=1+\mathscr{L}\left(\frac{d}{dt}J_0(|\sigma|\sqrt{t})\right)=1-\mathscr{L}\left(\frac{|\sigma|}{2}\frac{J_1(|\sigma|\sqrt{t})}{\sqrt{t}}\right).$$

Consequently

$$\frac{\sqrt{\pi}}{2\Gamma(m+\frac{1}{2})}p^m\tilde{H}(p)=\tilde{K}(p)-\frac{|\sigma|}{2}\tilde{K}(p)\cdot\mathscr{L}\left(\frac{J_1(|\sigma|\sqrt{t})}{\sqrt{t}}\right)(p).$$

Further by the equality

$$\frac{\sqrt{\pi}}{2\Gamma(m+\frac{1}{2})}\tilde{H}(p)=\tilde{K}(p)\frac{e^{-|\sigma|^2/4p}}{p^m}$$

it follows that $H^{(k)}(0)=\lim_{p\to+\infty}p^{k+1}\tilde{H}(p)=0$ for $k=0,1,\cdots,m-1$.

Now, inverting the Laplace transform we have

$$\frac{\sqrt{\pi}}{2\Gamma(m+\frac{1}{2})}H^{(m)}(r)=K(r)-\frac{|\sigma|}{2}\int_0^r K(s)\frac{J_1(|\sigma|\sqrt{r-s})}{\sqrt{r-s}}\,ds.$$

If we substitute $s=y^2$, $ds=2y\,dy$ and replace r by r^2 we have

$$\frac{\sqrt{\pi}}{2\Gamma(m+\frac{1}{2})}H^{(m)}(r^2)=K(r^2)-|\sigma|\int_0^r K(y^2)\frac{J_1(|\sigma|\sqrt{r^2-y^2})}{\sqrt{r^2-y^2}}y\,dy$$

whence we easily obtain (2). It also follows that $r^kG^{(k)}(r)$ is continuous on $[0, B)$ for $0\leqq k\leqq m$.

Next if $n=2m+1$ we have

$$\frac{\sqrt{\pi}}{2\Gamma(m+1)}p^m\tilde{H}(p)=\tilde{K}(p)\frac{e^{-|\sigma|^2/4p}}{p^{1/2}}.$$

Using that

$$\mathscr{L}^{-1}\left(\frac{e^{-|\sigma|^2/4p}}{p^{1/2}}\right)=\frac{1}{\sqrt{\pi}}\left(\frac{\cos(|\sigma|\sqrt{t})}{\sqrt{t}}\right)$$

and an argument similar to the preceding one we may obtain (3) and the statement about continuity for $r^kG^{(k)}(r)$. The proof of Theorem 5.1 is complete. We note that it is of course possible to obtain formulas for

$$\frac{d}{dr}\left(\frac{1}{2r}\frac{d}{dr}\right)^k(r^{n-1}G(r))\quad\text{for every } k \text{ with } 0\leqq k\leqq\left[\frac{n}{2}\right].$$

THEOREM 5.2. *If $g(x, r)=G(r)\,e^{i\langle\sigma,x\rangle}$ and if $r^kG^{(k)}(r)$ is continuous for $0\leqq r<B$ and $0\leqq k\leqq[n+1/2]$ then $g=Rf$ for some function $f(x, y)=F(y)\,e^{i\langle\sigma,x\rangle}$ where $F(y)$ is continuous for $0\leqq y<B$.*

For $n=2m$, $m\geqq 1$ we have

$$\frac{2\Gamma(m+\frac{1}{2})}{\sqrt{\pi}}F(y)=a(y)-|\sigma|y\int_0^y a(r)\frac{I_1(|\sigma|\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr \tag{7}$$

where

$$a(r)=\frac{d}{dr}\left(\frac{1}{2r}\frac{d}{dr}\right)^{m-1}(r^{2m-1}G(r)).$$

For $n=2m+1$, $m\geqq 0$ we have

$$F(y)=G(0)\cosh(|\sigma|y)+\frac{1}{\Gamma(m+1)}y\int_0^y b(r)\frac{\cosh(|\sigma|\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr \tag{8}$$

where

$$b(r)=\frac{d}{dr}\left(\frac{1}{2r}\frac{d}{dr}\right)^{m}(r^{2m}G(r)).$$

In particular for $n=1$

$$F(y)=G(0)\cosh(\sigma y)+y\int_0^y G'(r)\frac{\cosh(\sigma\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr \tag{9}$$

and for $n=2$,

$$F(y)=\frac{d}{dy}(yG(y))-|\sigma|y\int_0^y\frac{d}{dr}(rG(r))\frac{I_1(|\sigma|\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr. \tag{10}$$

Proof. Recalling the proof of Theorem 5.1 we see that to find a function $f(x, y)=F(y)\,e^{i\langle\sigma,x\rangle}$ satisfying $g=Rf$ is equivalent to solving (1) of Theorem 5.1 for F. Now this equation is, with the previous notation, equivalent to

$$\frac{2\Gamma((n+1)/2)}{\sqrt{\pi}}\tilde{K}(p)=p^{n/2}H(p)\,e^{|\sigma|^2/4p}.$$

If $n=2m$ then

$$\frac{2\Gamma(m+\frac{1}{2})}{\sqrt{\pi}}\tilde{K}(p)=p^{m}H(p)\,e^{|\sigma|^2/4p}.$$

Arguing as in the proof of Theorem 5.1, we obtain

$$\frac{2\Gamma(m+\frac{1}{2})}{\sqrt{\pi}}\tilde{K}(p)=p^{m}H(p)-p^{m}H(p)\mathcal{L}\left(\frac{|\sigma|}{2}\frac{I_1(|\sigma|\sqrt{t})}{\sqrt{t}}\right)$$

where we have used that

$$\mathcal{L}(I_0(|\sigma|\sqrt{t}))=\frac{e^{|\sigma|^2/4p}}{p},$$

$$e^{|\sigma|^2/4p}=1+\mathcal{L}\left(\frac{d}{dt}I_0(|\sigma|\sqrt{t})\right)$$

$$=1-\mathcal{L}\left(\frac{|\sigma|}{2}\frac{I_1(|\sigma|\sqrt{t})}{\sqrt{t}}\right).$$

Now $r^{m-1}G^{(m-1)}(r)$ is continuous on $[0, B)$ which implies that $H(0)=H'(0)=\cdots=H^{(m-1)}(0)=0$ and that $p^mH(p)=\mathcal{L}H^{(m)}(t)$. Inverting the Laplace transform, we easily

obtain (7) and the argument is complete for $n=2m$. Next, if $n=2m+1$ then

$$\frac{2\Gamma(m+1)}{\sqrt{\pi}}\tilde{K}(p)=p^{m+1}H(p)\frac{e^{|\sigma|^2/4p}}{p^{1/2}}.$$

Using that

$$\frac{e^{|\sigma|^2/4p}}{p^{1/2}}=\frac{1}{\sqrt{\pi}}\mathscr{L}\left(\frac{\cosh(|\sigma|\sqrt{t})}{\sqrt{t}}\right)$$

and that $H^{(m)}(0)=\Gamma(m+1)G(0)$ we obtain (8) after arguments similar to those previously used and the proof is finished.

THEOREM 5.3. *Let $g=Rf$ for some distribution $f\in\mathscr{S}'_e(\mathbb{R}^{n+1})$ and assume that the distribution g^* is for $0<r<B$ equal to a function $g^*(\sigma,r)$ such that $r^k(\partial/\partial r)^k g^*(\sigma,r)$ is continuous on $\mathbb{R}^n\times[0,B)\subset\mathbb{R}^{n+1}$ for $0\leqq k\leqq[(n+1)/2]$. Then the distribution f^* is on $\mathbb{R}^n\times[0,B)\subset\mathbb{R}^{n+1}$ equal to a continuous function $f^*(\sigma,y)$. Moreover, we have for $n=2m$ the inversion formula*

$$\frac{2\Gamma(m+\frac{1}{2})}{\sqrt{\pi}}f^*(\sigma,y)=a^*(\sigma,y)-|\sigma|y\int_0^y a^*(\sigma,r)\frac{I_1(|\sigma|\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr \tag{11}$$

where

$$a^*(\sigma,r)=\frac{\partial}{\partial r}\left(\frac{1}{2r}\frac{\partial}{\partial r}\right)^{m-1}(r^{2m-1}g^*(\sigma,r))$$

and for $n=2m+1$

$$f^*(\sigma,y)=g^*(\sigma,0)\cosh(|\sigma|y)+\frac{1}{\Gamma(m+1)}y\int_0^y b^*(\sigma,r)\frac{\cosh(|\sigma|\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr \tag{12}$$

where

$$b^*(\sigma,y)=\frac{\partial}{\partial r}\left(\frac{1}{2r}\frac{\partial}{\partial r}\right)^{m}(r^m g^*(\sigma,r)).$$

In particular

$$f^*(\sigma,y)=g^*(\sigma,0)\cosh(\sigma y)+y\int_0^y\frac{\partial g^*}{\partial r}(\sigma,r)\frac{\cosh(\sigma\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr$$

for $n=1$ and

$$f^*(\sigma,y)=\frac{\partial}{\partial y}(yg^*(\sigma,y))-|\sigma|y\int_0^y\frac{\partial}{\partial r}(rg^*(\sigma,r))\frac{I_1(|\sigma|\sqrt{y^2-r^2})}{\sqrt{y^2-r^2}}\,dr$$

for $n=2$.

Proof. Let us for $(\sigma,r)\subset\mathbb{R}^n\times[0,B)$ define a function $h^*(\sigma,r)$ by replacing f^* in (11) or (12) above, by h^*. Then h^* is continuous on $\mathbb{R}^n\times[0,B)$. Moreover, inverting these integral equations, we obtain

$$\left(\frac{|\sigma|}{2}\right)^{n-1}\frac{\sqrt{\pi}}{2\Gamma((n+1)/2)}g^*(\sigma,r)=\int_0^y h^*(\sigma,r)(r^2-y^2)^{(n-2)/4}J_{(n-2)/2}(|\sigma|\sqrt{r^2-y^2})\,dy$$

for $(\sigma,r)\in\mathbb{R}^n\times[0,B)$. Reverting the deduction in the proof of Theorem 5.1, we find that

$$\frac{\omega_n}{2}r^{n-1}g^*(\sigma,r)=\int_{|x|\leqq r}\frac{h^*(\sigma,\sqrt{r^2-|x|^2})}{\sqrt{r^2-|x|^2}}e^{i\langle\sigma,x\rangle}\,dx$$

for $(\sigma,r)\in\mathbb{R}^n\times[0,B)$.

Next suppose that $\varphi \in \mathcal{S}_r$. Then

$$\begin{aligned}(R^*\varphi)^*(\sigma, y) &= \int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\,dx \int_{\mathbb{R}^n} \varphi(z, \sqrt{|x-z|^2+y^2})\,dz \\ &= \int_{\mathbb{R}^n} e^{-i\langle\sigma,z\rangle}\,dz \int_{\mathbb{R}^n} e^{-i\langle\sigma,x-z\rangle}\varphi(z, \sqrt{|x-z|^2+y^2})\,dx \\ &= \int_{\mathbb{R}^n} e^{-i\langle\sigma,z\rangle}\,dz \int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\varphi(z, \sqrt{|x|^2+y^2})\,dx \\ &= \int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\,dx \int_{\mathbb{R}^n} e^{-i\langle\sigma,z\rangle}\varphi(z, \sqrt{|x|^2+y^2})\,dz \\ &= \int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\varphi^*(\sigma, \sqrt{|x|^2+y^2})\,dx.\end{aligned}$$

To proceed we introduce the following subspaces of $\mathcal{S}_e$ and $\mathcal{S}_r$.

$$\mathcal{S}_{e,B} = \{\psi^* \in \mathcal{S}_e\colon \psi^*(\sigma, y) = 0 \text{ if } |y| \geqq B\}$$

and

$$\mathcal{S}_{r,B} = \{\varphi \in \mathcal{S}_r\colon \varphi(x, r) = 0 \text{ for } r \geqq B\}.$$

Consider the mapping

$$R^*\colon \mathcal{S}_r \to \mathcal{S}_e$$

with

$$(R^*\varphi)^*(\sigma, y) = \int_{\mathbb{R}^n} e^{-i\langle\sigma,x\rangle}\varphi^*(\sigma, \sqrt{|x|^2+y^2})\,dx.$$

It is obvious that $R^*\varphi \in \mathcal{S}_{e,B}$ if $\varphi \in \mathcal{S}_{r,B}$.

LEMMA 5.4. $\overline{R^*(\mathcal{S}_{r,B})} = \mathcal{S}_{e,B}$.

Proof. Suppose, on the contrary, that $\overline{R^*(\mathcal{S}_{r,B})} \neq \mathcal{S}_{e,B}$. Then, according to the Hahn-Banach theorem (see [13, Chap. IV]) there exists some $f \in \mathcal{S}'_e$ and some $\psi \in \mathcal{S}_{e,B}$ such that $\langle f, R^*\varphi\rangle = 0$ for all $\varphi \in \mathcal{S}_{r,B}$ and so that $\langle f, \psi\rangle \neq 0$. Let $g = Rf$. Then $\langle g, \varphi\rangle = \langle f, R^*\varphi\rangle = 0$ for all $\varphi \in \mathcal{S}_{r,B}$. Consequently $g = 0$ for $r < B$ and by Theorem 4.1 we conclude that $f = 0$ for $|y| < B$. This contradicts that $\langle f, \psi\rangle \neq 0$ and the lemma is proved. Now suppose that $\varphi^* \in \mathcal{S}_{r,B}$. Then, using the previous relation between g^* and h^*,

$$\begin{aligned}\langle g^*, \varphi^*\rangle &= \omega_n \int_{\mathbb{R}^n} \int_0^\infty g^*(\sigma, r)\varphi^*(\sigma, r) r^n\,dr\,d\sigma \\ &= 2\int_{\mathbb{R}^n} \int_0^\infty \varphi^*(\sigma, r) r\,dr\,d\sigma \int_{|x|\leqq r} \frac{h^*(\sigma, \sqrt{r^2-|x|^2})}{\sqrt{r^2-|x|^2}}\, e^{i\langle\sigma,x\rangle}\,dx \\ &= 2\int_{\mathbb{R}^n} dx \int_{\mathbb{R}^n} e^{i\langle\sigma,x\rangle}\,d\sigma \int_{|x|}^\infty \frac{h^*(\sigma, \sqrt{r^2-|x|^2})}{\sqrt{r^2-|x|^2}}\varphi^*(\sigma, r) r\,dr \\ &= \{\text{substitute } r = \sqrt{|x|^2+y^2},\ y = \sqrt{r^2-|x|^2}\} \\ &= \int_{\mathbb{R}^n} dx \int_{\mathbb{R}^n} e^{i\langle\sigma,x\rangle}\,d\sigma \int_{\mathbb{R}} h^*(\sigma, y)\varphi^*(\sigma, \sqrt{|x|^2+y^2})\,dy \\ &= \int_{\mathbb{R}^{n+1}} h^*(\sigma, y)\,d\sigma\,dy \int_{\mathbb{R}^n} e^{i\langle\sigma,x\rangle}\varphi^*(\sigma, \sqrt{|x|^2+y^2})\,dx \\ &= \int_{\mathbb{R}^{n+1}} h^*(\sigma, y)(R^*\varphi)^*(\sigma, y)\,d\sigma\,dy.\end{aligned}$$

However, we also have $\langle g^*, \varphi^* \rangle = \langle f^*, (R^*\varphi)^* \rangle$ whence

$$\langle f^*, (R^*\varphi)^* \rangle = \int_{\mathbb{R}^{n+1}} h^*(\sigma, y)(R^*\varphi)^*(\sigma, y)\, d\sigma\, dy.$$

From the lemma it now follows easily that f^* is for $0 \leqq y < B$ equal to the function $h^*(\sigma, y)$. By the definition of h^* this completes the proof of Theorem 5.3.

We remark again that for $n = 1$ and 2 the direct formula and the inversion formula have the following form.

For $n = 1$

$$\frac{\pi}{2} g^*(\sigma, r) = \int_0^r f^*(\sigma, y) \frac{\cos(\sigma\sqrt{r^2 - y^2})}{\sqrt{r^2 - y^2}}\, dy,$$

$$f^*(\sigma, y) = g^*(\sigma, 0)\cosh(\sigma y) + y \int_0^y \frac{\partial g^*}{\partial r}(\sigma, r) \frac{\cosh(\sigma\sqrt{y^2 - r^2})}{\sqrt{y^2 - r^2}}\, dr.$$

For $n = 2$

$$rg^*(\sigma, r) = \int_0^r f^*(\sigma, y) J_0(|\sigma|\sqrt{r^2 - y^2})\, dy,$$

$$f^*(\sigma, y) = \frac{\partial}{\partial y}(yg^*(\sigma, y)) - |\sigma| y \int_0^y \frac{\partial}{\partial r}(rg^*(\sigma, r)) \frac{I_1(|\sigma|\sqrt{y^2 - r^2})}{\sqrt{y^2 - r^2}}\, dr.$$

COROLLARY 5.5. *Let $g = Rf$ for some distribution $f \in \mathcal{S}'_e(\mathbb{R}^{n+1})$ and assume that the distribution g^* is for $0 < r < B$ and $\sigma \in \Sigma$, where $\Sigma \subset \mathbb{R}^n$ is an open set, equal to a function $g^*(\sigma, r)$ such that $r^k(\partial/\partial r)^k g^*(\sigma, r)$ is continuous on $\Sigma \times [0, B) \subset \mathbb{R}^{n+1}$ for $0 \leqq k \leqq [(n+1)/2]$. Then the distribution f^* is on $\Sigma \times [0, B) \subset \mathbb{R}^{n+1}$ equal to a continuous function $f^*(\sigma, y)$ and the previously given inversion formulas are valid. In particular if $g^*(\sigma, r) = 0$ on $\Sigma \times [0, B)$ then so is $f^*(\sigma, y)$.*

The proof is omitted.

Finally we will demonstrate a theorem concerning the local range of the operator R. The proof of this theorem also suggests a procedure for the inversion of the operator R. However, for numerical purposes the given procedure may not be directly useful. For simplicity we formulate the theorem for $n = 2$.

THEOREM 5.6. *Let $n = 2$ and suppose that $g(x, r)$ and $rg'_r(x, r)$ are continuous in the set*

$$\{(x, r): 0 \leqq x_i \leqq a_i,\ i = 1, 2,\ 0 \leqq r \leqq B\} = K.$$

Then, given any $\varepsilon > 0$, there exists a function $h(x, r)$ such that

$$\sup_{(x,r) \in K} \{|g(x, r) - h(x, r)| + r|g'_r(x, r) - h'_r(x, r)|\} < \varepsilon$$

and such that $h = Rf$ for some function $f(x, y)$ which is continuous for $x \in \mathbb{R}^n$, $0 \leqq y \leqq B$.

Proof. First extend $g(x, r)$ to an even function in x by requiring that $g(\pm x_1, \pm x_2, r) = g(x_1, x_2, r)$. Then extend g periodically in x by the condition $g(x_1 + 2a_1, x_2, r) = g(x_1, x_2 + 2a_2, r) = g(x_1, x_2, r)$. Now we can find a function $h(x, r) = \sum g_\sigma(r)\, e^{i\langle \sigma, x \rangle}$ with the following properties. h is a trigonometric polynomial in (x_1, x_2) with the same periodicity as g. The coefficients $g_\sigma(r)$ are continuously differentiable for $0 \leqq r \leqq B$ and $\sup_{(x,r) \in K} (|g - h| + r|g'_r - h'_r|) < \varepsilon$. Now, in order to determine the

function f, we simply take $f(x, y) = \sum_\sigma f_\sigma(y)\, e^{i\langle\sigma, x\rangle}$ where

$$f_\sigma(y) = \frac{d}{dy}(yg_\sigma(y)) - |\sigma| y \int_0^y \frac{d}{dr}(rg_\sigma(r)) \frac{I_1(|\sigma|\sqrt{y^2 - r^2})}{\sqrt{y^2 - r^2}}\, dr.$$

Then according to Theorem 5.2 $h = Rf$ and the proof is finished.

In the previous proof we note that since $I_1(t)$ has exponential growth as $t \to \infty$ the formula for $f_\sigma(y)$ will give an amplification of terms g_σ with large frequencies $|\sigma|$. Although by choosing some limit for the bandwidth one may try to control the ill-posedness of the inversion problem, the required cut-off bandwidth will probably be so small that the method would not be directly useful.

6. Conclusions. In § 2 we have given a solution of the global inversion problem by means of Fourier transformation under very general and precise conditions on f and g. The inversion formula that is given is suitable for numerical purposes in many problems.

In § 3 we have a Sobolev estimate for the function f, and a reformulation of the form $f = c_n R^* K(Rf)$, where K is a certain pseudodifferential operator and R^* is the backprojection operator. This formula differs from the one given by Fawcett [3] which is of the form $f = c_n K R^* Rf$.

In § 4 we have, using the theory of § 2, extended a uniqueness theorem for the local problem to distributions.

In § 5 we gave inversion formulas for the local problem in terms of partial Fourier transforms with respect to the first n variables. These formulas are closely related to similar inversion formulas for the ordinary Radon transform, given in terms of integral equations involving Chebyshev polynomials of the first kind (see for instance the article by A. M. Cormack in [14]), and suffer from the same numerical limitations. Perhaps some modified version analogous to the one suggested in [14] might prove to be more efficient. Finally, considering Theorem 5.6, it is remarkable that the condition supp $\hat{g} \subset \{(\sigma, \rho)\colon \rho \geqq |\sigma|\}$ given in § 2 does not prevent the functions $g(x, r)$ from taking virtually arbitrary values in a bounded set $\{(x, r)\colon 0 \leqq x_i \leqq a_i,\ 0 \leqq r \leqq B\}$.

Acknowledgments. Dr. Hans Hellsten, National Defence Research Institute, Sweden, introduced me to the longwavelength SAR theory and to the problem of inverting spherical average for which I thank him. Anonymous referees have helped to improve the paper by helpful suggestions.

REFERENCES

[1] J. COHEN AND N. BLEISTEIN, *Velocity inversion procedure for acoustic waves*, Geophysics, 44 (1979), pp. 1077-1085.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. II, Wiley-Interscience, New York, 1962.

[3] J. A. FAWCETT, *Inversion of N-dimensional spherical means*, SIAM J. Appl. Math., 45 (1985), pp. 336-341.

[4] S. HELGASON, *The Radon Transform*, Birkhäuser, Boston, 1980.

[5] H. HELLSTEN AND L.-E. ANDERSSON, *An inverse method for the processing of synthetic aperture radar data*, Inverse Problems,, 4 (1987), pp. 111-124.

[6] H. HELLSTEN, J. KJELLGREN AND S. ÖDMAN, CARABAS-*large relative bandwidth* SAR *imagery*, internal report, 1986, National Defence Research Institute, FOA, Box 1165, S-581 11, Linköping, Sweden; IEEE Proc. Aerospace and Electronic Systems, submitted.

[7] M. HERBERTHSON, *A numerical implementation of an inversion formula for* CARABAS *raw data*, internal report D 30430-3.2, 1986, National Defence Research Institute, FOA, Box 1165, S-581 11, Linköping, Sweden.

[8] L. HÖRMANDER, *The analysis of Linear Partial Differential Operators* I, Springer-Verlag, Berlin, 1983.
[9] M. M. LAVRENTIEV, V. G. ROMANOV AND V. G. VASILIEV, *Multidimensional Inverse Problems for Differential Equations*, Lecture Notes in Mathematics 167, Springer-Verlag, Berlin, 1970.
[10] W. MAGNUS AND F. OBERHETTINGER, *Formulas and Theorems for the Functions of Mathematical Physics*, Chelsea Publishing, New York, 1949.
[11] V. G. ROMANOV, *Integral Geometry and Inverse Problems for Hyperbolic Equations*, Springer-Verlag, Berlin, 1968.
[12] E. STEIN, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[13] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1968.
[14] L. A. SHEPP, ED., *Computed tomography*, Lecture Notes Prepared for the American Mathematical Society Short Course, held in Cincinatti, Ohio, January 11-12, 1982.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

017

DEGREE OF APPROXIMATION OF REAL FUNCTIONS BY RECIPROCALS OF REAL AND COMPLEX POLYNOMIALS*

A. L. LEVIN† AND E. B. SAFF‡

Abstract. Let $E^c_{on}(f; I)(E^r_{on}(f; I))$ denote the error in best uniform approximation of a real continuous function f on a closed interval I by reciprocals of polynomials of degree $\leqq n$ with complex (real) coefficients. We investigate the rate at which $E^c_{on}(f; I)$ (or $E^r_{on}(f; I)$ provided $f \geqq 0$) can decrease. For example, we prove a Jackson type theorem and also present a class of functions for which reciprocal polynomial approximation is significantly better than polynomial approximation.

Key words. uniform approximation, reciprocals of polynomials, Jackson theorem

AMS(MOS) subject classifications. 41A20, 41A17

1. Introduction. For any real continuous function f on a closed interval I, let $E^r_{on}(f; I)$ and $E^c_{on}(f; I)$ denote the errors in best Chebyshev (uniform) approximation of f on I by reciprocals of polynomials of degree $\leqq n$ with real and complex coefficients respectively.

If f changes its sign on I, then obviously $E^r_{on}(f; I)$ does not approach zero as $n \to \infty$. On the other hand, a result of Walsh [8, Thm. IV] implies that any continuous function on I can be approximated arbitrarily close by reciprocals of *complex* polynomials; that is, $E^c_{on}(f; I) \to 0$. The aim of this paper is to investigate the rate at which $E^c_{on}(f; I)$ can decrease. For example, we prove a Jackson type theorem (Theorem 2.1) and also present a class of functions for which reciprocal polynomial approximation is significantly better than polynomial approximation (of the same degree).

Most of our results are formulated for the case $I = [-1, 1]$ but can be easily restated for an arbitrary finite interval. We also present some examples of approximation on the real line and on the unit disk.

The paper is organized as follows. In § 2 we state and discuss our main results. The proofs of these results are presented in §§ 3-6.

2. Main results. Our first result is the following Jackson type theorem.

THEOREM 2.1. *There exists an absolute constant M such that for any real $f \in C[-1, 1]$,*

$$E^c_{on}(f; [-1, 1]) \leqq M\omega(f; n^{-1}), \qquad n = 1, 2, 3, \cdots, \tag{2.1}$$

where $\omega(f; \delta)$ denotes the modulus of continuity of f on $[-1, 1]$.

Moreover, if f does not change its sign on $[-1, 1]$, then one can replace E^c_{on} by E^r_{on}:

$$E^r_{on}(f; [-1, 1]) \leqq M\omega(f; n^{-1}), \qquad n = 1, 2, 3, \cdots. \tag{2.2}$$

We remark that the estimate (2.1) follows from the estimate (2.2) and (via the usual Jackson theorem) from the following general result.

THEOREM 2.2. *For any real $f \in C[-1, 1]$,*

$$E^c_{0,3n}(f; [-1, 1]) \leqq 5(E^r_{on}(|f|; [-1, 1]) + E^r_{no}(f; [-1, 1])), \tag{2.3}$$

* Received by the editors May 27, 1986; accepted for publication (in revised form) February 12, 1987.

† Department of Mathematics, Everyman's University, Tel Aviv 61392, Israel. The research of this author was conducted while he was visiting the Institute for Constructive Mathematics at the University of South Florida, Tampa, Florida 33620.

‡ Institute for Constructive Mathematics, Department of Mathematics, University of South Florida, Tampa, Florida 33620. The research of this author was supported in part by the National Science Foundation.

where $E^r_{no}(f;[-1,1])$ *stands for the error in best Chebyshev approximation of* f *by polynomials of degree* $\leqq n$.

From Theorem 2.1 we obtain as a special case that

$$E^r_{on}(|x|^\alpha;[-1,1]) \leqq Mn^{-\alpha}, \qquad 0<\alpha\leqq 1. \tag{2.4}$$

For $0<\alpha<1$, this improves a result of Lungu [5]. For the case $\alpha=1$, the estimate (2.4) was proved by Newman and Reddy [6]. It turns out that the method of [6] can be modified to establish the estimate (2.4) for any $\alpha>0$. Since the matching lower bounds are also available (see Lungu [5]) we obtain the following result.

THEOREM 2.3. *For any* $\alpha>0$, *there exist positive constants* A_α, B_α *such that for any* $n=1,2,3,\cdots$, *the following hold*:

$$A_\alpha n^{-\alpha} \leqq E^c_{on}(|x|^\alpha;[-1,1]) \leqq E^r_{on}(|x|^\alpha;[-1,1]) \leqq B_\alpha n^{-\alpha}, \tag{2.5}$$

$$A_\alpha n^{-\alpha} \leqq E^c_{on}(|x|^\alpha \operatorname{sgn}(x);[-1,1]) \leqq B_\alpha n^{-\alpha}, \tag{2.6}$$

$$A_\alpha n^{-2\alpha} \leqq E^r_{on}(x^\alpha;[0,1]) \leqq B_\alpha n^{-2\alpha}. \tag{2.7}$$

Moreover, the constants A_α, B_α *may be written in the form* $A_\alpha=A^{-\alpha}$, $B_\alpha=C(B\alpha)^\alpha$, *where* A, B, C *are absolute constants* >1.

Note that the upper bound in (2.6) follows from that in (2.5) and (via Jackson's Theorem) from Theorem 2.2. The upper bound in (2.7) follows from that in (2.5) by the standard substitution $x\to x^2$.

The lower bounds in (2.5), (2.6) show that the estimate given in Theorem 2.1 is, in general, the least possible. Moreover, by considering the function $f(x)=x$, it is easy to see that no estimate of the kind $E^c_{on}(f;[-1,1])\leqq Mn^{-k}\omega(f^{(k)};n^{-1})$ (the analogue of Jackson's Theorem for differentiable functions) can be obtained. To get estimates better than $O(n^{-1})$ one has to make some assumptions concerning the zeros of f. The simplest theorem of this kind is the following one.

THEOREM 2.4. *Let* $f(\not\equiv 0)$ *be real-valued and analytic on* $[-1,1]$ *and assume* f *vanishes somewhere on* $[-1,1]$. *Denote by* r *the smallest order of the zeros of* f *in* $(-1,1)$ *and, by* s, *the smallest order of the zeros of* f *at* ±1 (*either* r *or* s *may be zero but not both*). *Define* k *by*

$$k:=\begin{cases} r & \text{if } s=0,\\ 2s & \text{if } r=0,\\ \min(r,2s) & \text{if } r>0,\quad s>0.\end{cases}$$

Then there exist positive constants $A(f)$, $B(f)$ *such that*

$$A(f)n^{-k} \leqq E^c_{on}(f;[-1,1]) \leqq B(f)n^{-k}, \qquad n=1,2,3,\cdots. \tag{2.8}$$

Moreover, the same estimates hold for $E^r_{on}(f;[-1,1])$ *provided* f *does not change sign on* $[-1,1]$.

Remark. The situation is more delicate if f is differentiable and does not change sign on $[-1,1]$. It may be true that in this case one can obtain the estimate $E^r_{on}(f;[-1,1])\leqq Mn^{-1}\omega(f';n^{-1})$ without any further assumptions on the structure of f. Even so, since $E^r_{on}(x^2;[-1,1])\neq 0$, no further refinement involving the modulus of continuity of higher derivatives is possible.

Our next result exhibits a class of functions that can be approximated by reciprocals of polynomials much better than by polynomials (of the same degree). The common feature of functions of this class is that they vanish on a set of intervals but not at

isolated points. To demonstrate why such functions are "well approximable" by reciprocals of polynomials, consider the following example. Let

$$x_+ := \begin{cases} x, & x \geqq 0, \\ 0, & x < 0. \end{cases}$$

Then, from the well-known degree of polynomial approximation to $|x|$, we have $E_{no}(x_+; [-1, 1]) \geqq Cn^{-1}$. On the other hand, we can find a real polynomial $p_n(x)$ such that (see (2.7)) $|x - 1/p_n(x)| \leqq An^{-2}$ for $0 \leqq x \leqq 1$. In particular, $p_n(0) \geqq A^{-1}n^2$. It can be shown that $p_n(x)$ is monotonic on $(-\infty, 0)$ and consequently $p_n(x) \geqq A^{-1}n^2$ for $x < 0$. It follows that

$$|x_+ - 1/p_n(x)| = 1/p_n(x) < An^{-2} \quad \text{for } x < 0,$$

and we obtain

$$E^r_{on}(x_+; [-1, 1]) \leqq An^{-2}.$$

We now formulate the general result.

THEOREM 2.5. *Let $[a_j, b_j], j = 1, 2, \cdots, N$ ($N \geqq 1$), be mutually disjoint subintervals of $[-1, 1]$. For each j, let $f_j \in C^3[a_j, b_j]$ be real-valued and satisfy $f_j(a_j) = f_j(b_j) = 0$, $f_j'(a_j + 0) \neq 0$, $f_j'(b_j - 0) \neq 0$, and assume that $f_j \neq 0$ in (a_j, b_j). Define the function f on $[-1, 1]$ by*

$$f(x) := \begin{cases} f_j(x) & a_j \leqq x \leqq b_j, \quad j = 1, 2, \cdots, N, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E^c_{on}(f; [-1, 1]) \leqq A(f)n^{-2}, \qquad n = 1, 2, 3, \cdots. \tag{2.9}$$

Further, if f does not change sign on $[-1, 1]$ then $E^r_{on}(f; [-1, 1]) \leqq A(f)n^{-2}$.

Notice that if $N \geqq 2$ or if $N = 1$ and $[a_1, b_1] \neq [-1, 1]$, then f is not differentiable somewhere in $(-1, 1)$ and consequently $E_{no}(f; [-1, 1]) \neq O(n^{-1-\varepsilon})$ for any $\varepsilon > 0$.

So far we have discussed the direct theorems. What about inverse results? To have any chance of proving that f is differentiable on $[-1, 1]$, we have to assume at least (in view of Theorem 2.5) that $E^c_{on}(f; [-1, 1]) = o(n^{-2})$. Under some additional assumptions on the behavior of f near its zeros we can then prove the differentiability of f. For example, if f is piecewise continuously differentiable on $[-1, 1]$ and satisfies $E^c_{on}(f; [-1, 1]) = o(n^{-2})$, then f is continuously differentiable on $[-1, 1]$. At present, the proper formulation of an inverse theorem for differentiable functions is not clear. We confine ourselves to the following Bernstein type result, which was essentially proved by J. L. Walsh [8].

THEOREM 2.6. *For any complex-valued function $f(\not\equiv 0)$ on $[-1, 1]$, the following conditions are equivalent:*

(i) $\limsup_{n\to\infty} \{E^c_{on}(f; [-1, 1])\}^{1/n} < 1$.

(ii) *f is analytic on $[-1, 1]$ and does not vanish there.*

Our final result deals with approximation on the real axis.

THEOREM 2.7. *Let $K(x)$, $L(x)$ be real polynomials of degrees k and l respectively, with $k \leqq l - 1$. If $L(x) \neq 0$ for x real, then*

$$E^r_{on}(|K(x)|/L(x); \mathbb{R}) = O(n^{(k/l)-1}), \tag{2.10}$$

$$E^c_{on}(K(x)/L(x); \mathbb{R}) = O(n^{(k/l)-1}). \tag{2.11}$$

Furthermore, if $K(x)$ does not change sign for $x \in \mathbb{R}$ then

$$E^r_{on}(K(x)/L(x); \mathbb{R}) = O(n^{(2k/l)-2}). \tag{2.12}$$

For the special case $K(x)=x$, $L(x)=1+x^{2m}$, the estimates (2.10), (2.12) were proved by Newman and Reddy [6]. The lower bounds obtained in [6] show that the estimates of Theorem 2.7 are, in general, the best possible.

3. Jackson type theorems.

Proof of Theorem 2.1. To prove the estimate (2.1) it suffices to prove the corresponding estimate for the case of approximation of 2π-periodic functions on the interval $[-\pi, \pi]$ by reciprocals of trigonometric polynomials of degree n. In what follows we use the notation and the estimates that appear in the book of Lorentz [4, p. 55-56].

For any 2π-periodic function g, let

$$J_n(g; x) := \int_{-\pi}^{\pi} g(x+t)K_n(t)\,dt$$

be the Jackson operator. Since

$$\int_{-\pi}^{\pi} K_n(t)\,dt = 1, \quad \int_{-\pi}^{\pi} |t|^k K_n(t)\,dt = O(n^{-k}), \quad k = 1, 2,$$

we obtain that

$$\int_{-\pi}^{\pi} |g(x+t)-g(x)|K_n(t)\,dt \leqq c_1\omega(g; n^{-1}) \tag{3.1}$$

and that

$$\int_{-\pi}^{\pi} |g(x+t)-g(x)|^2 K_n(t)\,dt \leqq c_2[\omega(g; n^{-1})]^2, \tag{3.2}$$

where c_1, c_2 are absolute constants.

Consider now the function

$$f_\varepsilon(x) := f(x) + i\varepsilon, \tag{3.3}$$

where f is a given real 2π-periodic function and $\varepsilon > 0$ will be chosen later. Since f is real, $1/f_\varepsilon$ is continuous on $[-\pi, \pi]$. Furthermore,

$$\omega(f_\varepsilon; n^{-1}) = \omega(f; n^{-1}) \tag{3.4}$$

and

$$|1/f_\varepsilon(x)| \leqq 1/\varepsilon, \qquad -\pi \leqq x \leqq \pi. \tag{3.5}$$

Define the trigonometric polynomial p_n of degree $\leqq n$ by

$$p_n(x) := J_n(1/f_\varepsilon; x).$$

Then

$$\begin{aligned} |1/f_\varepsilon(x) - p_n(x)| &\leqq \int_{-\pi}^{\pi} |1/f_\varepsilon(x) - 1/f_\varepsilon(x+t)|K_n(t)\,dt \\ &= \int_{-\pi}^{\pi} \frac{|f_\varepsilon(x+t)-f_\varepsilon(x)|}{|f_\varepsilon(x)f_\varepsilon(x+t)|} K_n(t)\,dt. \end{aligned}$$

Hence,

$$|1-f_\varepsilon(x)p_n(x)| \leqq \int_{-\pi}^{\pi} |f_\varepsilon(x+t)-f_\varepsilon(x)| \frac{1}{|f_\varepsilon(x+t)|} K_n(t)\,dt$$

$$\leqq \frac{1}{\varepsilon} c_1\omega(f_\varepsilon; n^{-1}) = \frac{1}{\varepsilon} c_1\omega(f; n^{-1})$$

by (3.1), (3.4) and (3.5).

The choice

$$\varepsilon = 2c_1\omega(f; n^{-1}) \tag{3.6}$$

therefore yields

$$|f_\varepsilon(x)p_n(x)| \geqq \tfrac{1}{2}, \qquad -\pi \leqq x \leqq \pi. \tag{3.7}$$

In particular, $p_n \neq 0$ on $[-\pi, \pi]$. Now

$$|f_\varepsilon(x)-1/p_n(x)| = |1/f_\varepsilon(x)-p_n(x)| \cdot |f_\varepsilon(x)/p_n(x)|$$

$$\leqq \int_{-\pi}^{\pi} \frac{|f_\varepsilon(x+t)-f_\varepsilon(x)|}{|f_\varepsilon(x)f_\varepsilon(x+t)|} \cdot \left|\frac{f_\varepsilon(x)}{p_n(x)}\right| \cdot K_n(t)\,dt$$

$$\leqq 2\int_{-\pi}^{\pi} |f_\varepsilon(x+t)-f_\varepsilon(x)| \cdot \left|\frac{f_\varepsilon(x)}{f_\varepsilon(x+t)}\right| K_n(t)\,dt \quad \text{(by (3.7))}$$

$$\leqq 2\int_{-\pi}^{\pi} |f_\varepsilon(x+t)-f_\varepsilon(x)| K_n(t)\,dt$$

$$+2\int_{-\pi}^{\pi} \frac{|f_\varepsilon(x+t)-f_\varepsilon(x)|^2}{|f_\varepsilon(x+t)|} K_n(t)\,dt$$

$$\leqq 2c_1\omega(f_\varepsilon; n^{-1}) + \frac{2}{\varepsilon}\int_{-\pi}^{\pi} |f_\varepsilon(x+t)-f_\varepsilon(x)|^2 K_n(t)\,dt$$

by (3.1) and (3.5). Thus, from (3.2), (3.4) and (3.6) we deduce that

$$|f_\varepsilon(x)-1/p_n(x)| \leqq (2c_1+c_2/c_1)\omega(f; n^{-1}).$$

This yields (see (3.3), (3.6)) the first part of Theorem 2.1.

For the second part, we suppose that $f \geqq 0$ on $[-\pi, \pi]$ and set

$$f_\varepsilon(x) := f(x)+\varepsilon. \tag{3.3'}$$

Then the polynomial $J_n(1/f_\varepsilon; x)$ will have real coefficients. The rest of the proof remains the same. □

Remark. Although it does not follow immediately from the above argument, Theorem 2.1 holds, more generally, for any *complex-valued* continuous function f on $[-1, 1]$. The proof of this fact will appear in [2]. Moreover, for a special class of functions f, our methods can be adapted to obtain a Jackson-type theorem for approximation by reciprocal polynomials on the unit disk $|z| \leqq 1$. For example, in [3] we prove that

$$E^c_{on}((z-1)^\alpha; |z| \leqq 1) \leqq Mn^{-\alpha}, \qquad 0 < \alpha \leqq 1.$$

We now proceed to the proof of Theorem 2.2, which uses an idea of Trefethen [7].

LEMMA 3.1. *Let p_n be a real polynomial of degree $\leqq n$. Then*

$$E^c_{0,3n}(p_n; I) \leqq 4E^r_{on}(|p_n|; I).$$

Proof. Let q_n be a real polynomial of degree $\leqq n$ satisfying

$$\max_{x\in I}\big||p_n(x)|-1/q_n(x)\big| = E^r_{on}(|p_n|;\, I) =: \varepsilon.$$

Then

$$|p_n^2(x)-1/q_n^2(x)| \leqq \varepsilon(2|p_n(x)|+\varepsilon), \qquad x\in I.$$

Define the complex polynomial Q_{3n} of degree $\leqq 3n$ by

$$Q_{3n}(x) := (p_n(x)-i\varepsilon)q_n^2(x).$$

Then

$$\begin{aligned}|p_n(x)-1/Q_{3n}(x)| &\leqq \left|p_n(x)-\frac{p_n^2(x)}{p_n(x)-i\varepsilon}\right| \\ &\quad+\left|\frac{p_n^2(x)}{p_n(x)-i\varepsilon}-\frac{1}{(p_n(x)-i\varepsilon)q_n^2(x)}\right| \\ &\leqq \varepsilon\left|\frac{ip_n(x)}{p_n(x)-i\varepsilon}\right|+\varepsilon\frac{2|p_n(x)|+\varepsilon}{|p_n(x)-i\varepsilon|}\leqq 4\varepsilon,\end{aligned}$$

since p_n is real. □

Proof of Theorem 2.2. Let p_n be any real polynomial of degree $\leqq n$. With obvious simplification of notation we obtain the following:

$$\begin{aligned}E^c_{0,3n}(f) &\leqq \|f-p_n\|+E^c_{0,3n}(p_n) \\ &\leqq \|f-p_n\|+4E^r_{on}(|p_n|) \quad \text{(by Lemma 3.1)} \\ &\leqq \|f-p_n\|+4[E^r_{on}(|f|)+\|\,|f|-|p_n|\,\|] \\ &\leqq 5\|f-p_n\|+4E^r_{on}(|f|).\end{aligned}$$

Hence, on choosing p_n such that $\|f-p_n\| = E^r_{no}(f)$, Theorem 2.2 follows. □

4. Approximation of powers of x. The lower bounds for $E^r_{on}(|x|^\alpha;[-1,1])$ and for $E^r_{on}(x^\alpha;[0,1])$ were proved (for $0<\alpha\leqq 1$) by Lungu [5]. The proof for other cases is exactly the same. For the proof of the upper bounds it suffices to show (as we mentioned in the Introduction) that

$$E^r_{on}(|x|^\alpha;[-1,1]) \leqq B_\alpha n^{-\alpha}, \qquad \alpha>0. \tag{4.1}$$

Following an idea in Newman and Reddy [6], we consider the kernel

$$\varphi_n(t) := t^{\alpha-1}\left(\frac{T_n(t)}{t}\right)^{2k} \tag{4.2}$$

where n is odd, k is the smallest integer satisfying $k\geqq\alpha$ and T_n denotes the nth degree Chebyshev polynomial of the first kind. Define

$$p(x) := \frac{1}{Cx^\alpha}\int_0^x \varphi_n(t)\,dt, \qquad x>0 \tag{4.3}$$

where $C := \int_0^1 \varphi_n(t)\,dt$.

Clearly, $p(x)$ is an even polynomial of degree $2k(n-1)$. By evenness we consider only $x\in[0,1]$. Write

$$\frac{1}{p(x)}-x^\alpha = \frac{x^\alpha\int_x^1\varphi_n(t)\,dt}{\int_0^x\varphi_n(t)\,dt}. \tag{4.4}$$

As in [6] we make use of the estimates $|T_n(t)/t| \leqq n$, $|T_n(t)/t| \leqq 1/t$ for $0 < t \leqq 1$ and $|T_n(t)/t| \geqq 2n/\pi$ for $0 \leqq t \leqq \sin(\pi/2n)$. It follows that

$$\varphi_n(t) \leqq n^{2k} t^{\alpha-1}, \quad \varphi_n(t) \leqq t^{\alpha-1-2k} \quad \text{for } 0 < t \leqq 1,$$

and

$$\varphi_n(t) \geqq \left(\frac{2}{\pi}\right)^{2k} n^{2k} t^{\alpha-1} \quad \text{for } 0 < t \leqq \sin\left(\frac{\pi}{2n}\right).$$

We consider now two cases.

Case 1. Suppose $0 \leqq x \leqq \sin(\pi/2n)$. In this case, we have

$$\int_x^1 \varphi_n(t)\,dt \leqq \int_0^{1/n} + \int_{1/n}^1 \leqq \int_0^{1/n} n^{2k} t^{\alpha-1}\,dt + \int_{1/n}^1 t^{\alpha-1-2k}\,dt$$

$$\leqq \frac{1}{\alpha} n^{2k-\alpha} + \frac{1}{2k-\alpha} n^{2k-\alpha}$$

$$\leqq \frac{2}{\alpha} n^{2k-\alpha} \quad \text{since } k \geqq \alpha.$$

Also,

$$\int_0^x \varphi_n(t)\,dt \geqq \int_0^x \left(\frac{2}{\pi}\right)^{2k} n^{2k} t^{\alpha-1}\,dt = \left(\frac{2}{\pi}\right)^{2k} n^{2k} \cdot \frac{1}{\alpha} x^{\alpha}.$$

It now follows from (4.4) that

$$0 < \frac{1}{p(x)} - x^{\alpha} \leqq 2\left(\frac{\pi}{2}\right)^{2k} n^{-\alpha}, \qquad 0 \leqq x \leqq \sin\left(\frac{\pi}{2n}\right). \tag{4.5}$$

Case 2. Suppose $\sin(\pi/2n) \leqq x \leqq 1$. In this case, we have

$$\int_x^1 \varphi_n(t)\,dt \leqq \int_x^1 t^{\alpha-1-2k}\,dt < \int_x^{\infty} t^{\alpha-1-2k}\,dt$$

$$= \frac{1}{2k-\alpha} x^{\alpha-2k} \leqq \frac{1}{\alpha} x^{\alpha-2k} \quad \text{since } k \geqq \alpha.$$

Hence,

$$x^{\alpha} \int_x^1 \varphi_n(t)\,dt \leqq \frac{1}{\alpha} x^{2(\alpha-k)} \leqq \frac{1}{\alpha} \left(\sin\frac{\pi}{2n}\right)^{2(\alpha-k)} \leqq \frac{1}{\alpha} n^{2(k-\alpha)}.$$

Also,

$$\int_0^x \varphi_n(t)\,dt \geqq \int_0^{\sin(\pi/2n)} \varphi_n(t)\,dt \geqq \int_0^{\sin(\pi/2n)} \left(\frac{2}{\pi}\right)^{2k} n^{2k} t^{\alpha-1}\,dt$$

$$= \left(\frac{2}{\pi}\right)^{2k} n^{2k} \frac{1}{\alpha} \left(\sin\frac{\pi}{2n}\right)^{\alpha} \geqq \frac{1}{\alpha} \left(\frac{2}{\pi}\right)^{2k} n^{2k-\alpha}.$$

It follows that

$$0 < \frac{1}{p(x)} - x^{\alpha} \leqq \left(\frac{\pi}{2}\right)^{2k} n^{-\alpha}, \qquad \sin\left(\frac{\pi}{2n}\right) \leqq x \leqq 1. \tag{4.6}$$

From (4.5) and (4.6) it follows that

$$E^r_{0,2k(n-1)}(|x|^{\alpha}; [-1, 1]) \leqq 2(\pi/2)^{2k} n^{-\alpha}$$

(recall that $p(x)$ is of degree $2k(n-1)$). Using a standard technique, the last inequality implies (4.1) with a constant B_α of the form $C(B\alpha)^\alpha$, where $B, C>1$ are absolute constants. Analyzing the proof of lower bounds given in Lungu [5], we see that A_α may be taken of the form $A^{-\alpha}$, where $A>1$ is an absolute constant. The proof of Theorem 2.3 is complete. □

An appropriate change of variable yields the following corollary.

COROLLARY 4.1. *For any $a\in[-1,1]$ and for any $\alpha>0$ there exist constants c_1, c_2 (depending on a, α) such that*

$$c_1 n^{-\alpha} \leqq E^r_{on}(|x-a|^\alpha;[-1,1]) \leqq c_2 n^{-\alpha} \quad \text{if } |a|<1, \tag{4.7}$$

$$c_1 n^{-2\alpha} \leqq E^r_{on}(|x-a|^\alpha;[-1,1]) \leqq c_2 n^{-2\alpha} \quad \text{if } |a|=1. \tag{4.8}$$

The same estimates hold for $E^c_{on}(|x-a|^\alpha \operatorname{sgn}(x-a);[-1,1])$.

We conclude this section with a simple lemma. This lemma together with Corollary 4.1 enable us to obtain upper bounds for $E_{on}(f;[-1,1])$, where f is a finite product of functions of type $|x-a|^\alpha$ or $|x-a|^\alpha \operatorname{sgn}(x-a)$.

LEMMA 4.2. *For any complex-valued continuous functions f, g, on I, there is a constant K (independent of n) such that*

$$E_{0,2n}(fg;I) \leqq K(E_{on}(f;I)+E_{on}(g;I)), \tag{4.9}$$

where E_{om} stands for E^r_{om} or for E^c_{om}.

Proof. Choose the polynomials p_n, q_n such that

$$\left\| f-\frac{1}{p_n}\right\| = E_{on}(f;I), \qquad \left\| g-\frac{1}{q_n}\right\| = E_{on}(g;I),$$

where $\|\cdot\|$ denotes the uniform norm on I. Since

$$\begin{aligned}\left\| fg-\frac{1}{p_nq_n}\right\| &= \left\| \left(f-\frac{1}{p_n}\right)g+\frac{1}{p_n}\left(g-\frac{1}{q_n}\right)\right\| \\ &\leqq \|g\|E_{on}(f;I)+2\|f\|E_{on}(g,I),\end{aligned}$$

the result follows. □

5. Well-approximable functions (Proof of Theorem 2.5). The proof of Theorem 2.5 given in this section will be split into several lemmas. We shall use the following notation:

$$f(x)_+ = \begin{cases} f(x) & \text{if } f(x)\geqq 0, \\ 0 & \text{otherwise.}\end{cases}$$

LEMMA 5.1. *For any $\alpha>0$ and for any $n=1,2,3,\cdots$,*

$$E^r_{on}(x_+^\alpha;[-1,1]) \leqq E^r_{on}(x_+^\alpha;(-\infty,1]) \leqq B_\alpha n^{-2\alpha}. \tag{5.1}$$

Proof. We consider the kernel $\varphi_n(t)$ and the polynomial $p(x)$ as in the proof of Theorem 2.3 (see formulas (4.2), (4.3)) with α replaced by 2α. From the proof of Theorem 2.3 we obtain that

$$\left| x^{2\alpha}-\frac{1}{p(x)}\right| \leqq B_\alpha n^{-2\alpha}, \qquad 0\leqq x\leqq 1.$$

Recall that $p(x)$ is an even polynomial of degree $2k(n-1)$, where k is the smallest integer satisfying $k\geqq 2\alpha$. Define the polynomial $Q(x)$ by $Q(x):=p(\sqrt{x})$. Then

$$|x^\alpha - 1/Q(x)| \leqq B_\alpha n^{-2\alpha}, \qquad 0\leqq x\leqq 1. \tag{5.2}$$

Since $T_n(t)/t$ has the form $(-1)^{(n-1)/2}\sum_{j=0}^{(n-1)/2}(-1)^j a_j t^{2j}$, $a_j>0$, the polynomial $[T_n(t)/t]^{2k}$ has a similar form (except that the factor preceding the summation is now 1). Hence the polynomial $p(x)$ is of the form $\sum_{j=0}^{k(n-1)}(-1)^j b_j x^{2j}$, $b_j>0$, and therefore $Q(x)$ has the form

$$Q(x)=\sum_{j=0}^{k(n-1)}(-1)^j b_j x^j, \qquad b_j>0.$$

It follows that $Q(x)>Q(0)$ for x negative and we obtain from (5.2) that

$$0<1/Q(x)<1/Q(0)<B_\alpha n^{-2\alpha}, \qquad -\infty<x<0.$$

Hence,

$$|x_+^\alpha-1/Q(x)|\leqq B_\alpha n^{-2\alpha}, \qquad -\infty<x\leqq 1. \qquad \square$$

LEMMA 5.2. *Let p be a real polynomial. Then*

$$E_{on}^r(p_+;[-1,1])\leqq cn^{-2}\|p_+\|(\deg p)^2, \tag{5.3}$$

where c is an absolute constant and $\|\cdot\|$ denotes the uniform norm on $[-1,1]$.

Proof. By the proof of Lemma 5.1, there exists a polynomial $q_m(x)$ of degree m such that

$$|x_+-1/q_m(x)|\leqq cm^{-2}, \qquad -\infty<x\leqq 1. \tag{5.4}$$

Let $\deg p=:k$ and define the polynomial $Q_{mk}(x)$ of degree mk by

$$Q_{mk}(x)=\frac{1}{\|p_+\|}q_m(p(x)/\|p_+\|).$$

The substitution $x\to p(x)/\|p_+\|$ in (5.4) yields:

$$\|p_+(x)-1/Q_{mk}(x)\|\leqq cm^{-2}\|p_+\|=c(mk)^{-2}\|p_+\|k^2.$$

Hence the lemma is established for n of the form mk. The result for arbitrary n follows by a standard technique. $\square$

LEMMA 5.3. *Let f be a nonvanishing real continuous function on $[-1,1]$. Then*

$$c_1E_{no}^r(1/f;[-1,1])\leqq E_{on}^r(f;[-1,1])\leqq c_2E_{no}^r(1/f;[-1,1]), \tag{5.5}$$

where $c_1>0$, $c_2>0$ depend on f.

The proof of Lemma 5.3 is straightforward.

LEMMA 5.4. *For any $0<a<1$, there is a constant c (depending on a) such that*

$$E_{on}^r((|x|-a)_+;[-1,1])\leqq cn^{-2}. \tag{5.6}$$

Proof. It suffices to prove that

$$E_{on}^r((\sqrt{x}-a)_+;[0,1])\leqq cn^{-2}.$$

To show this write

$$(\sqrt{x}-a)_+=(x-a^2)_+\frac{1}{\sqrt{x}+a}.$$

From Lemma 5.1 it follows (by linear transformation of the variable) that

$$E_{on}^r((x-a^2)_+;[0,1])\leqq cn^{-2}. \tag{5.7}$$

Further, we can extend the function $1/(\sqrt{x}+a)$, $x \geqq a^2$, to the interval $[0, 1]$ in such a way that the resulting function, $g(x)$ say, will belong to $C^2[0, 1]$ and will be positive on $[0, 1]$. By Lemma 5.3 and by Jackson's Theorem for differentiable functions (see e.g. [4, p. 57]) we obtain that

$$E^r_{on}(g; [0, 1]) \leqq cn^{-2}. \tag{5.8}$$

Since $(x-a^2)_+ g(x) = (\sqrt{x}-a)_+$ on $[0, 1]$, the inequalities (5.7), (5.8) and Lemma 4.2 yield the desired estimate. □

LEMMA 5.5. *For any $0<a<1$ there is a constant c (depending on a) such that*

$$E^c_{on}((|x|-a)_+ \operatorname{sgn}(x); [-1, 1]) \leqq cn^{-2}. \tag{5.9}$$

Proof. Write

$$(|x|-a)_+ \operatorname{sgn}(x) = x^3(|x|-a)_+|x|^{-3},$$

and extend the function $|x|^{-3}$, $|x| \geqq a$, to the interval $[-1, 1]$ as a twice differentiable positive function. The lemma now follows (as in the proof of Lemma 5.4) from Theorem 2.3, Lemma 4.2 and Lemma 5.4. □

Using the proof similar to that of Lemma 5.2 we obtain the following.

LEMMA 5.6. *Let $p(x)$ be a real polynomial and let $0<a<\|p\|$, where $\|\cdot\|$ denotes the uniform norm on $[-1, 1]$. Then*

$$E^c_{on}((|p(x)|-a)_+ \operatorname{sgn} p(x); [-1, 1]) \leqq cn^{-2}\|p\|(\deg p)^2, \tag{5.10}$$

where c depends only on a.

Proof of Theorem 2.5. We first consider the case when all functions f_j are of the same sign (positive, say). Define the polynomial $p(x)$ by

$$p(x) := -\prod_{j=1}^{N} (x-a_j)(x-b_j).$$

Then $p(x)>0$ on each interval (a_j, b_j). It follows that the function

$$g_j(x) := f_j(x)/p(x), \qquad x \in [a_j, b_j],$$

is positive on $[a_j, b_j]$ and belongs to $C^2[a_j, b_j]$. We can find now a function $G(x) \in C^2[-1, 1]$ that is positive on $[-1, 1]$ and coincides with g_j on $[a_j, b_j]$, $j=1, 2, \cdots, N$. Since $p(x) \leqq 0$ whenever $f(x)=0$, we can write

$$f(x) = p_+(x)G(x).$$

By Lemma 5.2,

$$E^r_{on}(p_+; [-1, 1]) \leqq cn^{-2}$$

(c depends on f) and by Lemma 5.3

$$E^r_{on}(G; [-1, 1]) \leqq cn^{-2},$$

since $1/G$ is twice differentiable. Applying Lemma 4.2 we obtain that

$$E^r_{on}(f; [-1, 1]) \leqq c(f)n^{-2}.$$

For the general case, when the f_j are of arbitrary signs, define the function φ on $\bigcup_{j=1}^{N} [a_j, b_j]$ by

$$\varphi(x) := \begin{cases} f_j(x)+\frac{1}{2} & \text{if } f_j>0, \\ f_j(x)-\frac{1}{2} & \text{if } f_j<0. \end{cases}$$

Then

$$(|\varphi(x)|-\tfrac{1}{2}) \operatorname{sgn} \varphi(x) = f_j(x), \qquad x \in [a_j, b_j], \quad j=1, 2, \cdots, N. \tag{5.11}$$

Next, we claim that there is a polynomial $P(x)$ of some fixed but large degree, such that

(i) $|P(x)|<\frac{1}{2}$ for $x\in[-1,1]\setminus\bigcup_{j=1}^{N}[a_j,b_j]$,

(ii) $|P(x)|>\frac{1}{2}$ for $x\in(a_j,b_j)$, $j=1,2,\cdots N$,

(iii) $P(a_j)=P(b_j)=\begin{cases}\frac{1}{2} & \text{if } f_j>0,\\ -\frac{1}{2} & \text{if } f_j<0.\end{cases}$

This can be seen as follows. The function φ satisfies conditions (ii), (iii). Extend it to $[-1,1]$ in such a way that it will satisfy (i) and will belong to $C^3[-1,1]$. Now approximate φ simultaneously with φ' by a polynomial P that interpolates φ, φ' at a_j, b_j, $j=1,2,\cdots,N$. If the degree of P is large enough, the norms $\|\varphi-P\|$, $\|\varphi'-P'\|$ will be arbitrarily small (see Chalmers and Taylor [1, pp. 55-56]). From this it follows easily that P will satisfy (i)-(iii).

From this construction we obtain that the function

$$g(x)=\frac{(|\varphi(x)|-\frac{1}{2})\operatorname{sgn}\varphi(x)}{(|P(x)|-\frac{1}{2})\operatorname{sgn}P(x)}$$

is positive on $[a_j,b_j]$, $j=1,2,\cdots,N$ and has there two continuous derivatives. Extend g to $[-1,1]$ preserving its sign and the differentiability. From (5.11) and from the definition of g we obtain that

$$[(|P(x)|-\tfrac{1}{2})_+\operatorname{sgn}P(x)]g(x)=f(x),\qquad -1\leqq x\leqq 1.$$

By Lemma 5.3 and Jackson's Theorem,

$$E^c_{on}(g;[-1,1])\leqq cn^{-2}.$$

Also, by Lemma 5.6,

$$E^c_{on}((|P(x)|-\tfrac{1}{2})_+\operatorname{sgn}P(x);[-1,1])\leqq cn^{-2}.$$

Finally, Lemma 4.2 yields

$$E^c_{on}(f;[-1,1])\leqq cn^{-2}.$$ □

6. Approximation of analytic functions (Proofs of Theorems 2.4 and 2.6).

Proof of Theorem 2.4. If $f(\not\equiv 0)$ is real analytic on $[-1,1]$, we can write

$$f(x)=(x+1)^{a_1}(x-1)^{a_2}\prod_{j=1}^{N}(x-x_j)^{b_j}\cdot g(x), \tag{6.1}$$

where $a_1,a_2,b_1,\cdots,b_N$ are nonnegative integers, $|x_j|<1$ for $j=1,2,\cdots,N$ and g is real analytic and nonvanishing on $[-1,1]$. By Corollary 4.1, we have $E^r_{on}((x+1)^{a_1};[-1,1])\leqq cn^{-2a_1}$, $E^r_{on}((x-1)^{a_2};[-1,1])\leqq cn^{-2a_2}$, $E^c_{on}((x-x_j)^{b_j};[-1,1])\leqq cn^{-b_j}$. Also, by Lemma 5.3 and by Bernstein's Theorem (cf. [4, p. 76]),

$$\limsup_{n\to\infty}[E^c_{on}(g;[-1,1])]^{1/n}<1.$$

Applying Lemma 4.2 we obtain the estimate

$$E^c_{on}(f;[-1,1])\leqq cn^{-k},$$

where k is defined in Theorem 2.4.

For the lower bound in (2.8) we write $f(x)=(x-x_j)^{b_j}\varphi_j(x)$, where $\varphi_j(x_j)\neq 0$ and apply the argument in Lungu [5] to obtain $E^c_{on}(f;[-1,1])\geqq cn^{-b_j}$, $j=1,2,\cdots,N$. We omit the details. □

Concerning Theorem 2.6, Walsh [8] proves the corresponding result for approximation on a Jordan region. He asserts that the result is also true for Jordan arcs, but does not provide the proof. For completeness we provide the details.

Proof of Theorem 2.6. The implication (ii)$\Rightarrow$(i) is trivial (apply Lemma 5.3 and Bernstein's Theorem). Assume now that

$$\limsup_{n\to\infty} (E_{on}^{c}(f;[-1,1]))^{1/n} = q < 1,$$

and let $P_n(x)$, $n = 1, 2, 3, \cdots$, be polynomials for which

$$\|f - 1/P_n\| = E_{on}^{c}(f;[-1,1]).$$

It suffices to prove that $f \neq 0$ on $[-1,1]$, since then Lemma 5.3 and Bernstein's Theorem will imply the analyticity of f on $[-1,1]$. Suppose that f vanishes somewhere on $[-1,1]$. Since $f \not\equiv 0$, we can find an interval $I \subset [-1,1]$ such that $f \neq 0$ inside I but vanishes at one of its endpoints. Assume, for simplicity, that $I = [-\delta, \delta]$, $\delta < 1$, and $f(\delta) = 0$. Then

$$\limsup_{n\to\infty} |1/P_n(\delta)|^{1/n} \leqq q,$$

which implies that

$$\liminf_{n\to\infty} |P_n(\delta)|^{1/n} \geqq 1/q. \tag{6.2}$$

Pick $\delta_1 < \delta$. Since $f \neq 0$ on $[-\delta_1, \delta_1]$, there exists a constant $M = M(\delta_1)$ such that $|P_n(x)| \leqq M$ for $n = 1, 2, 3, \cdots$, and for $x \in [-\delta_1, \delta_1]$. Then (cf. [4, p. 43])

$$|P_n(\delta)| \leqq M\left(\frac{1+\sqrt{1-(\delta_1/\delta)^2}}{\delta_1/\delta}\right)^n.$$

It follows that

$$\liminf_{n\to\infty} |P_n(\delta)|^{1/n} \leqq \frac{1+\sqrt{1-(\delta_1/\delta)^2}}{\delta_1/\delta} < \frac{1}{q},$$

provided δ_1 is close enough to δ. This contradicts (6.2). □

7. Approximation on the real line.

Proof of Theorem 2.7. Consider the polynomial $p(x)$ defined by formulae (4.2) and (4.3) with $\alpha = 1$, $k = 1$. By the proof in § 4, we obtain that p is a polynomial of degree $2(n-1)$ satisfying

$$||x| - 1/p(x)| \leqq An^{-1} \quad \text{for } -1 \leqq x \leqq 1. \tag{7.1}$$

From (4.4) we also obtain that

$$0 < |x| - 1/p(x) < |x| \quad \text{for } |x| > 1. \tag{7.2}$$

For $T > 0$, set

$$F_T := \{x \in \mathbb{R}: |K(x)| \leqq \|K\|_{[-T,T]}\}.$$

Then we obtain by the substitution $x \to |K(x)|/\|K\|_{[-T,T]}$ in (7.1), (7.2) that

$$\left||K(x)| - \frac{1}{q(x)}\right| \leqq \begin{cases} An^{-1}\|K\|_{[-T,T]} & \text{if } x \in F_T, \\ |K(x)| & \text{if } x \in \mathbb{R}\setminus F_T, \end{cases} \tag{7.3}$$

where $q(x) := \|K\|_{[-T,T]}^{-1} p(K(x)/\|K\|_{[-T,T]})$ is a polynomial of degree $\leqq 2k(n-1)$. Dividing (7.3) by $L(x)$ and setting $a := \min_{\mathbb{R}} |L(x)| > 0$, we obtain that

$$\left| \frac{|K(x)|}{L(x)} - \frac{1}{r(x)} \right| \leqq \begin{cases} Aa^{-1}n^{-1}\|K\|_{[-T,T]} & \text{if } x \in F_T, \\ |K(x)|/L(x) & \text{if } x \in \mathbb{R}\setminus F_T, \end{cases}$$

where $r(x) := q(x)L(x)$ is a polynomial of degree $2k(n-1)+l$. For T large enough we have $\|K\|_{[-T,T]} = O(T^k)$ and $|K(x)/L(x)| = O(T^{k-l})$ uniformly for $x \in \mathbb{R}\setminus F_T$ and consequently

$$E^r_{0,2k(n-1)+l}(|K(x)|/L(x); \mathbb{R}) = O(T^k)n^{-1} + O(T^{k-l}).$$

When we choose $T = n^{1/l}$ we obtain the first assertion of Theorem 2.7.

For the third assertion we assume that $K(x) \geqq 0$, $L(x) > 0$ on $\mathbb{R}$ and make use of the polynomial $\tilde{p}(x) := p(\sqrt{x})$, where p is defined by (4.2) and (4.3) with $\alpha = 2$, $k = 2$. Then

$$|x - 1/p(x)| \leqq \begin{cases} An^{-2} & \text{if } 0 \leqq x \leqq 1, \\ x & \text{if } x > 1, \end{cases}$$

and we can repeat the above proof choosing eventually $T = n^{2/l}$.

It remains to prove the second assertion of Theorem 2.7 (formula (2.11)). Let p be the polynomial satisfying (7.1)-(7.2). Using the method of the proof of Lemma 3.1 we see that when we choose

$$\tilde{p}(x) = (x - i\varepsilon)p^2(x), \qquad \varepsilon := E^r_{on}(|x|; [-1, 1]), \tag{7.4}$$

we obtain

$$|x - 1/\tilde{p}(x)| \leqq An^{-1} \quad \text{for } |x| \leqq 1. \tag{7.5}$$

For $|x| > 1$ we have

$$|1/\tilde{p}(x)| = \frac{1}{|x - i\varepsilon|} \frac{1}{p^2(x)} \leqq \frac{|x|^2}{|x - i\varepsilon|} \leqq |x| \quad \text{(by (7.2))}.$$

Hence

$$|x - 1/\tilde{p}(x)| \leqq 2|x| \quad \text{for } |x| > 1. \tag{7.6}$$

Using (7.5) and (7.6) the proof can be completed as above. □

REFERENCES

[1] B.L. CHALMERS AND G. D. TAYLOR, *Uniform approximation with constraints*, Jber. d. Dtsch. Math. Verein, 81 (1979), pp. 49-86.

[2] A. L. LEVIN AND E. B. SAFF, *Jackson type theorems in approximation by reciprocals of polynomials*, Rocky Mountain J., to appear.

[3] ———, *Some examples in approximation on the unit disk by reciprocals of polynomials*, in Tampa Approximation Seminar Proceedings, Lecture Notes in Math., Springer-Verlag, Berlin, to appear.

[4] G. G. LORENTZ, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[5] K. N. LUNGU, *Best approximation of $|x|$ by rational functions of the form $1/P_n(x)$*, Siberian Math. J., 15 (1974), pp. 1152-1156.

[6] D. J. NEWMAN AND A. R. REDDY, *Rational approximation to $|x|/(1+x^{2m})$ on $(-\infty, \infty)$*, J. Approx. Theory, 19 (1977), pp. 231-238.

[7] L. N. TREFETHEN, Personal communication.

[8] J. L. WALSH, *On approximation to an analytic function by rational functions of best approximation*, Math. Z., 38 (1934), pp. 163-176.

SIAM J. MATH. ANAL.
Vol. 19, No. 1, January 1988

018

THE NEWTONIAN GRAPH OF A COMPLEX POLYNOMIAL*

MICHAEL SHUB†, DAVID TISCHLER‡ AND ROBERT F. WILLIAMS§

Abstract. In a recent paper [4] Smale posed as an important problem in complexity theory, characterization of the graph G_f of the Newtonian vector field N_f for a complex polynomial f. Such graphs are known to be connected and acyclic and Smale conjectured that these two properties completely characterize them. The purpose of this paper is to prove this conjecture, after adding an additional hypothesis (part 3 of the definition of "dynamic graph," § 2). In additon we give an example and prove a proposition to show this is necessary (see "Counterexamples" in § 2).

We present the proof as Theorem C in § 5 using the topological characterization of analytic maps given by Stöilow [5] in 1929. Bill Thurston pointed us in this direction, though considering the fact that G. T. Whyburn was the major professor of one of us, we should not have needed this help. In addition we present direct proofs of three special cases as Theorem A, Theorem B and Example 7. While this was being written an independent proof was given in the generic case (Theorem A) in [2].

Sections 1 and 2 are devoted to basic properties and to a list of examples designed to acquaint the reader (and the writers) with various aspects of Newtonian Graphs.

Key words. polynomial, root, Newtonian vector field, Newtonian Graph

AMS(MOS) subject classifications. 58F, 65H, 68Q

1. The Newtonian vector field. Given a smooth map $f: \mathbf{R}^n \to \mathbf{R}^n$ the Newtonian vector field for f, N_f, is defined by

$$N_f(x) = -(Df_x)^{-1}(f(x)).$$

Let $\phi_t: \mathbf{R}^n \to \mathbf{R}^n$ be the corresponding flow. Then a computation carried out below using the chain rule shows that $f(\phi_t(x)) = e^{-t}f(x)$. Thus f maps orbits of ϕ_t into rays pointing toward 0. This is essentially the geometric content to Newton's method for seeking zeros of f.

Alternatively, one defines

$$V_f(x) = -\tfrac{1}{2}\operatorname{grad} \|f(x)\|^2 = -(Df_x)^{\mathrm{tr}} f(x).$$

For conformal maps, and in particular analytic maps, of one complex variable (by the Cauchy-Riemann equations) the inverse of a matrix and its transpose differ only by a scalar multiple. Therefore, the vector fields V_f and N_f also differ only by a scalar multiple, except where $(Df)^{-1}$ is undefined.

We next collect some facts for f a polynomial of one complex variable:

1. N_f or V_f have the same solution curves as

(a) $-f(z)\bar{f}'(z)$ or

(b) $-\sum (z-a_j)/|z-a_j|^2$, $\{a_j\}$ the zeros of f.

Thus the field is the sum of forces toward a_j, each inversely proportional to the distance from a_j.

To see (a),

$$N_f(z) = -\frac{f(z)}{f'(z)} \cdot \frac{\bar{f}'(z)}{\bar{f}'(z)} = -(1/|f'(z)|^2) f(z)\bar{f}'(z)$$

* Received by the editors July 8, 1985; accepted for publication (in revised form) December 1, 1986. This work was partially supported by National Science Foundation grants in Mathematics and Computer Science and the National Science Foundation U.S.-Latin America Cooperative Science Program.

† IBM, Thomas J. Watson Research Center, Yorktown Heights, New York 10598-0218.

‡ Queens College and the Graduate School and University Center, City University of New York, New York, New York 10036-8099.

§ Department of Mathematics, Northwestern University, Evanston, Illinois 60201.

so that N_f differs from $-f(z)\bar{f}'(z)$ only by the scalar function $|f'(z)|^{-2}$. Note however that this scalarization converts poles of N_f to zeros of V_f.

(b)
$$-f/f' = -[(\log f)']^{-1} = -\left[\sum \frac{1}{z-a_i}\right]^{-1}$$
$$= -\left[\sum \frac{\overline{z-a_i}}{|z-a_i|^2}\right]^{-1} = -\frac{\bar{A}}{\bar{A}}\left[\sum \frac{\overline{z-a_i}}{|z-a_i|^2}\right]^{-1}$$
$$= \left(-\sum \frac{z-a_i}{|z-a_i|^2}\right)(1/|A|^2),$$

where $A = \sum \overline{z-a_i}/|z-a_i|^2$.

2. Properties of N_f and V_f:

(a) $f(\phi_t(x)) = e^{-t}f(x)$

(b) N_f and V_f have attractors (sinks) at the zeros a_j of f.

(c) The only other rest points of V_f are the zeros θ_j of f':

(i) generic zeros of f' are hyperbolic saddles of V_f;

(ii) at multiple zeros of f' V_f has multipronged saddles ("monkey saddles" and worse);

(iii) For θ a zero of f', the orbits leaving θ, called together the unstable manifold $W^u(\theta)$, consist of n algebraic curves emanating from θ at equal angles because $f(z) = c_1 + c_2(z-\theta)^n +$ higher order terms, $c_2 \neq 0$.

(d) Multiple zeros a of f have no geometric effect on the corresponding sinks. Only the *velocity* of the flow toward a is increased.

(e) (Gauss-Lucas Theorem). The convex hull of the sinks $\{a_j\}$ contains the saddles $\{\theta_j\}$. The flow of N_f or V_f is inwardly transverse to any convex curve containing the $\{a_j\}$.

Proof of 2. To prove (a) we compute

$$\frac{d}{dt} f(\phi_t(x)) = Df(\phi_t(x)) \cdot \frac{d}{dt}\phi_t(x)$$
$$= Df(\phi_t(x)) \cdot (-Df(\phi_t(x)))^{-1} \cdot f(\phi_t(x))$$
$$= -f(\phi_t(x)).$$

Thus $f(\phi_t(x)) = \rho_t$ where

$$\frac{d\rho}{dt} = -\rho \quad \text{at } t = 0, \quad \rho = f(x)$$

which has $e^{-t}f(x)$ as solutions.

Parts (b) and (d) follow from the "attracting force" version of N_f, $-\sum(z-a_j)/|z-a_j|^2$. Similarly, if all the a_j are on the side of a line, the vector field is transverse to this line, which proves the second part of (e). The first part follows from the second part.

The form $-f(z)\bar{f}'(z)$ shows that the zeros of f and f' are the only rest points. Part (a) implies (c)(ii)-(c)(iii); the inverse image of a directed line under the map $z \to c_1 + c_2(z-\theta)^n +$ h.o.t. consists of $2n$ directed curves pointing alternatively toward and away from θ, with tangents evenly spaced at θ.

In fact, property 2(a) is proven for general $C^1 f$ and 2(b) is true for simple zeros of $C^1 f$ since at such points the derivative of N_f is $-I$.

When $f'(\theta)=0$, $DV_f(\theta)(w)=-\overline{f''(\theta)w}f(\theta)$. If $f(\theta)\neq 0$ and $f''(\theta)\neq 0$ then the 2×2 real matrix representing this linear transformation has strictly negative determinant and trace 0. Thus the eigenvalues of $DV_f(\theta)$ are real and have opposite sign.

We have borrowed here from [3] and [1] where a desingularization of Newton's method is discussed in more variables as well.

2. Newtonian graphs and special terminology. Let f be a complex polynomial with $\{a_j\}$ its zeros and $\{\theta_k\}$ the zeros of f' which are not zeros of f. Let $V_f=-f(z)\overline{f'(z)}$ be the gradient vector field of f as in § 1. Let $W^u(\theta_k)$ be the "unstable manifold" of θ_j, i.e., the union of all solutions which limit on θ_k as $t\to-\infty$. Note that they in turn limit on some a_j or some other θ_k, as $t\to+\infty$. Define

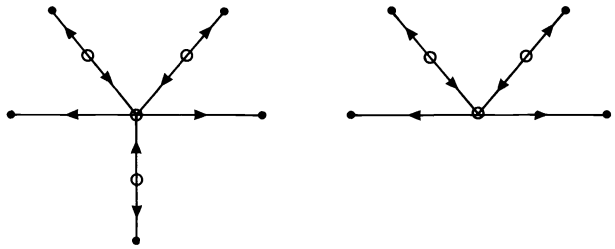
$$G_f=\{a_j\}\smile\{\theta_j\}\smile W^u(\theta_j).$$

Then G_f is a finite graph with distinguished vertices a_j, θ_k and directed "edges" $W^u(\theta_k)$.

For nonrepeated zeros θ_k of f', $W^u(\theta_k)\smile\{\theta_k\}$ is a smooth curve but in degenerated cases $W^u(\theta_k)$ consists of 3 or more prongs.

The sinks a_j of G_f have weights α_j where α_j is the multiplicity of a_j.

Counterexamples. CE1 below is not homeomorphic to any Newtonian graph G_f, f a polynomial. CE2 is not isotopic to any such G_f.



CE1 CE2

These facts follow from the general proposition.

PROPOSITION. *For f any complex polynomial and v a vertex of G_f at most one incoming edge can lie between any two outgoing edges.*

Proof. Suppose the contrary. Then choose a small circle J around v and note that it passes through points A, B, C, D where

(i) $A<B<C<D$ on J;

(ii) there is no other point of $J\frown$ (outgoing edge) between A and D on J;

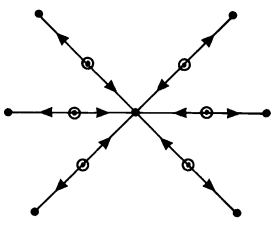
(iii) A and D lie on rays beginning at v;

(iv) B and C lie on rays terminating at v.

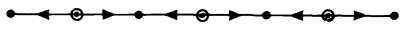
Then f has a singularity at v (of index $\geqq 2$) and f maps the open arc (B, C) of J completely around $f(v)$ since it intersects the ray $\{mf(v)\,|\,m \text{ real } m>1\}$ in the two points $f(B), f(C)$. $f|(B, C)$ does not intersect the ray $\{mf(v)\,|\,m \text{ real}, 0<m<1\}$ which is a contradiction.

Examples. In our sketches we indicate the weights only if different from 1.

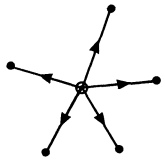
1. z^n-z $n=7$



2. Real roots

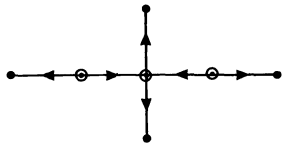


3. $z^d - 1$ $d = 5$

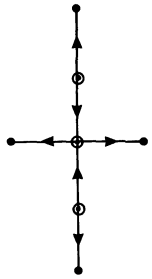


multiple saddle or
unstable star with 5 prongs.

4. $(z^2 + a^2)(z^2 - b^2)$ saddle connections

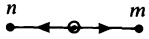


$a < b$



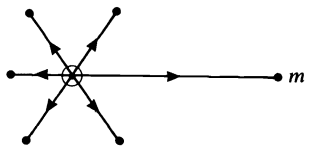
$a > b$

5. $z^n(z-c)^m$



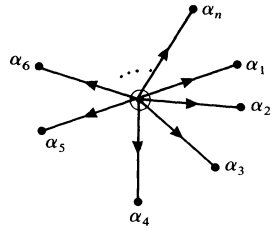
sinks of weights n and m.

6. There is a polynomial P, monic and of degree n such that $((z-c)^m P)' = (n+m)(z-c)^{m-1} z^n$ m, n positive integers. (This is proved in Lemma 4.1; we sketch the graph for such a P, where $n = 5$ next.)



$n = 5$

7. Given positive integers $\alpha_1, \cdots, \alpha_n$ there is a polynomial of degree $\alpha_1 + \cdots + \alpha_n$ with graph

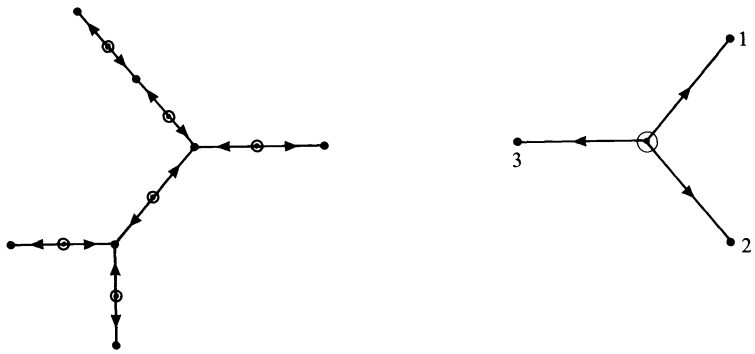


unstable star with weights

Remark. For any complex polynomial f, the graph G_f is connected and acyclic.

Proof. A cycle in the graph, G_f, would bound a finite region in the plane, but this contradicts 2(a). The flow is a gradient flow with ∞ as the only source. Thus the plane is the union of G_f and $W^u(\infty)$, the unstable manifold of ∞. Thus G_f has the homotopy type of the plane (or a point) and so G_f is connected.

Remark. Any connected acyclic finite graph can be embedded in the plane, and sometimes in distinct-*nonisotopic* ways. Embeddings $f: G \to \mathbf{R}^2$ and $g: G \to \mathbf{R}^2$ are *isotopic* provided there is a continuous 1-parameter family f_t, $0 \leqq t \leqq 1$ such that $f_0 = f$, $f_1 = g$ and f_t is an embedding of G for each t, $0 \leqq t \leqq 1$. This last property, that each f_t be an embedding, distinguishes isotopy from homotopy; were it to be dropped, one could reverse the orientation of the next two examples, by pushing one of the legs through another one. The following two examples have two isotopy classes of embedding determined by the cyclic order at the circled vertices.



PROPOSITION. *Generically V_f is structurally stable, having*

(a) *hyperbolic saddles,*

(b) *no saddle connections, and*

(c) *all weights* $= 1$.

Proof. Generically f and f' have no repeated zeros which proves (a) and (c). Any one saddle connection can be removed by an arbitrarily small perturbation of one of the vertices involved, noting for example that a saddle connection implies that two critical values lie on the same ray. Thus proceeding one at a time, one can remove all saddle connections by a perturbation so small as not to effect those already broken. Structural stability now follows from Peixoto's theorem (see Palis-de Melo [8]).

In order for an abstract finite acyclic graph to be the Newtonian graph of a complex polynomial, it must have certain properties best described in dynamical terms.

By a *dynamic graph* we mean a finite directed graph with two types of vertices, which we call *saddles* and *sinks* subject to the following conditions:

1) At a sink, all edges are directed inward (i.e., toward the sink).

2) Saddles have at least two outwardly directed edges.

3) At a saddle, any two adjacent outwardly directed edges have at most one inwardly directed edge between them. (Each such edge must then connect two saddles, and is thus called a *saddle connection.*)

4) Each sink has a weight which is a positive integer, often 1. The weights have no effect on the geometry of a dynamic graph.

It follows that a dynamic graph falls into natural units each consisting of a saddle together with all edges directed away from it. These will be called *unstable stars* or *k-prongs* where k is the number of issuing edges, $k = 2, 3, \cdots$. A 2-prong is also called a *hyperbolic saddle.*

3. The generic case. Our first theorem is a special case of each of the two others, but its proof is easy. In addition, while this was being written, other authors [2] have found a proof independently of this part of our results.

THEOREM A. *Given an acyclic dynamic graph with all saddles hyperbolic, no saddle connections and all weights* $= 1$, *there is a complex polynomial* f *whose graph* G_f *is isotopic to* G.

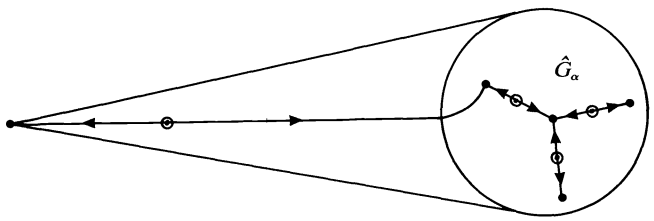
Proof. First, there exist sinks $v_0, v_1, \cdots, v_m \in G$, and subgraphs $G_\alpha \subset G$ such that for each α, $v_\alpha \in G_\alpha$ and $G_{\alpha+1}$ consists of G_α together with one additional hyperbolic saddle having v_α as one of its 2 sinks.

We proceed by induction on α. The graph $\{v_0\}$ is realized by the polynomial z. Thus assume we have a generic polynomial P_α such that the graph $\hat{G}_\alpha$ of P_α is isotopic to G_α. Let D be a (round) disk containing the zeros of P_α. Then $D \supset \hat{G}_\alpha$ and the field V_{P_α} is inwardly transverse to the boundary ∂D of D.

Let $f(z) = P_\alpha(\lambda(z-c))$ where $|\lambda| = 1$ and c is real. Note that the graph of f is conjugate to $\hat{G}_\alpha$ by a linear conjugacy: $f'(z) = \lambda P'_\alpha(\lambda(z-c))$ so that $V_f(z) = -\bar{\lambda} P_\alpha(\lambda(z-c))\bar{P}'_\alpha(\lambda(z-c)) = \bar{\lambda} V_{P_\alpha}(\lambda(z-c))$ which gives $V_f(z) = \lambda^{-1} V_{P_\alpha}(\lambda(z-c))$.

Let $g(z) = zf(z)$. We claim that for c sufficiently large and certain λ, $|\lambda| = 1$, the graph of g is isotopic to $G_{\alpha+1}$. In fact the field V_g on D differs from V_f only by the summand $-1/z$ and this is essentially constant $= -1/c$. Thus for $c \gg 1$, the part of the graph of g related to the zeros and saddles of f is isotopic to $\hat{G}_\alpha$ by structural stability.

Note that the "ice-cream cone" region sketched below contains the zeros of g and hence the dynamics of G_g. It follows that the graph G_g of g consists of $\hat{G}_\alpha$ together with a single edge added and that for c large, the saddle of this new edge is outside D.



Now as λ varies in the unit circle, the disk D and the dynamics of V_{P_α} rotates through a full circle as well, by structural stability. Thus our new edge arrives at any one of the entering orbits, for the appropriate choice of λ. This is important below, but here we have more leeway, because there is an open set of orbits limiting on the

sink v_α, and have the correct deployment with respect to $\hat{G}_\alpha$. Thus for some choice of λ the graph G_g of g is isotopic to $G_{\alpha+1}$.

This completes the inductive step and the proof of Theorem A.

4. Nongeneric saddles and Theorem B. The purpose of this section is to prove Thoerem B. This uses the fact that our example 6 is the Newtonian graph of a complex polynomial which we prove as Lemma 4.1. Example 6 is used in the proof of Theorem B (implicitly) and in example 7, below.

LEMMA 4.1. *Given integers m, n and $c \in \mathbf{C}$, there exists a monic polynomial P of degree $n+m$ such that $P(0) \neq 0$ for $c \neq 0$ and $((z-c)^m P(z))' = (n+m)z^n(z-c)^{m-1}$.*

Proof. This linear ODE reduces to $(z-c)P' + mP = (n+m)z^n$. We try a solution of the form $P = z^n + \sum_{i=0}^{n-1} b_i z^i$ and note that the coefficient of z^n is $n+m$ on both sides. Proceeding downward, for $j = n-1,\ n-2, \cdots, 0$ one has the formulas $jb_j - c(j+1)b_{j+1} + mb_j = 0$ for the coefficient of z^j. Thus

$$b_j = \frac{c(j+1)}{m+j} b_{j+1}$$

gives our solution, which terminates with b_0 as $b_{-1} = 0$.

THEOREM B. *Given an oriented acyclic embedded dynamic graph G with no saddle connections and all weights 1, there exists a polynomial f with graph G_f isotopic to G.*

Proof. We proceed as before with v_α, $G_\alpha \subset G$ by induction on α. Here, however, $G_{\alpha+1}$ is G_α with a *multiple saddle* attached at $v_\alpha \in G_\alpha$. Suppose then that P_α is a polynomial yielding the graph G_α and that D is a disk around 0 containing all G_α (as above) and w, the field of P_α transverse to ∂D. Let $f(z) = P_\alpha(\lambda(z-c))$. For simplicity we assume 0 is a zero of P_α and set $f(z) = P_\alpha(\lambda(z-c))$.

Now $G_{\alpha+1}$ is G_α with a saddle edge added at the vertex v_α. Say the new unstable star has $(n+1)$ exiting edges, for some $n \geqq 1$. Then we want our next polynomial g to have derivative $z^n f'(z)$, or $g(z) = \int_c^z w^n f'(w)\,dw$, some $c \in \mathbf{C}$ which we choose to be real. Here we use the same c as that in $f(z) = P_\alpha(\lambda(z-c))$. Then integrating by parts,

$$g(z) = z^n f(z) - n \int_c^z w^{n-1} f(w)\, dw,$$

as $f(c) = 0$.

Then the gradient field for g is given by

$$V_g = -|z|^{2n} f(z)\bar{f}'(z) + n\bar{z}^n \bar{f}'(z) \int_c^z w^{n-1} f(w)\, dw.$$

Now near c we scalarize to $V_g/|z|^{2n}$ which gives

$$-f(z)\bar{f}'(z) + \frac{n\bar{f}'(z)}{z^n} \int_c^z w^{n-1} f(w)\, dw,$$

so that we have our given field with an error term,

$$E = \frac{n\bar{f}'(z)}{z^n} \int_c^z w^{n-1} f(w)\, dw.$$

Then one can estimate E on D_c by

$$|E| \leqq \frac{n(c+\delta)^{n-1} |f'(z) f(\zeta)| \delta}{(c-\delta)^n} \quad \text{some } z, \zeta \in D_c$$

which goes to 0 for large c, where $\delta =$ diameter of D.

Next to check the C^1 part of the error, we note that E is differentiable as a map from $\mathbf{R}^2$ to $\mathbf{R}^2$, as conjugation is real analytic. Hence

$$E' = \frac{nz^{n-1}\bar{f}'(z)f(z)}{z^n} + \frac{n\bar{f}''(w)\int_c^z w^{n-1}f(w)\,dw}{z^n} - \frac{n^2\bar{f}'(z)\int_c^z w^{n-1}f(w)\,dw}{z^{n+1}}.$$

These clearly go to zero on D_c as $c \to \infty$. Thus the field of g on the disk is near that $f(\lambda(c-z))$ and has exactly the same saddles. It follows that for c large enough, the graph G_g is isotopic to G_α with an unstable star of $n+1$ prongs added.

But just as in the proof of Theorem A, we can choose λ so that this last edge is added in the correct "angle."

This completes the proof of Theorem B.

PROPOSITION (*Example* 7). *For any (integral) weights $\alpha_1, \cdots, \alpha_k$ there is an unstable star with these weights. Equivalently, there is a solution $a_1, a_2, \cdots, a_k$, $a_i \neq 0$ of the equation*

$$\left(\prod_{n=1}^{k}(z-a_n)^{\alpha_n}\right)' = (\alpha_1 + \cdots + \alpha_k)z^{k-1}\prod_{n=1}^{k}(z-a_n)^{\alpha_n - 1}.$$

Furthermore, any cyclic ordering of the weights (see the above remark) can be realized.

Proof. This ODE leads to an unstable star with weight α_i at a_i, provided it has a solution with the a_i distinct. To solve it is equivalent to solving a set of $k-1$ equations in k unknowns, which we augment by one equation.

$$(j)\quad \sum_{n=1}^{k}\alpha_n \sum_{\substack{i_1<i_2<\cdots<i_j \\ i_\alpha \neq i_n}} a_{i_1}a_{i_2}\cdots a_{i_j} = 0, \qquad j = 1, 2, \cdots, k-1;$$

$$a_k - 1 = 0.$$

Now the function $F_\alpha : \mathbf{C}^k \to \mathbf{C}^k$ defined by the left-hand sides, where $\alpha = \alpha_1, \cdots, \alpha_k$ is a proper map and has a positive Jacobian (see [1, p. 294]). Thus F_α can be extended to the $2k$-sphere $\mathbf{C}^k \cup \{\infty\}$ so that it has a solution $a_1, \cdots, a_k$.

In fact, by counting degrees we see that there are $(k-1)!$ solutions. To see that these contain all of the isotopy classes we need a homotopy argument. First, we may as well suppose that the α_n's are numbered in their desired cyclic order. Next let $\alpha(t)$ be a path, $0 \leqq t \leqq 1$, where

$$\alpha(0) = (1, 1, \cdots, 1) \quad \text{and}$$

$$\alpha(1) = (\alpha_1, \alpha_2, \cdots, \alpha_k) \quad \text{and}$$

$$\alpha_i(t) \neq \alpha_j(t) \quad \text{for } i \neq j \text{ and } 0 < t < 1.$$

Now by Lemma 4.1, above, there is a singular graph of weights $1, 1, \cdots, 1$, that is, the regular unstable star of k petals, and thus there is a solution $a_n(0)$ $n = 1, \cdots, k$ of the equation $F_{\alpha(0)} = 0$. Then $F_{\alpha(t)}$ is an analytic isotopy of this algebraic function, so there is a unique arc of solutions $\{a_n(t)\}$ to the equations $F_{\alpha(t)} = 0$. This gives a 1-parameter path of functions $f_t(z) = \prod_1^k (z-a_n)^{a_n}$, and Newtonian graphs G_t, whose end G_1 has its weights in the correct order as G_t is an isotopy of graphs. The weights $\{\alpha_n(t)\}$ do vary with t, of course, but G_t is an isotopy of the graphs if we disregard weights. One could also interpret G_t as an isotopy of weighted graphs. The $f_t(z)$ are proper and their Newtonian graphs are acyclic, connected, etc.

5. The general case via Stöilow's theorem.

THEOREM C. *Given an acyclic dynamic graph $G \subset \mathbf{R}^2$ there is a polynomial f such that the Newton graph G_f is isotopic to f.*

DEFINITIONS (see Whyburn [7]). A map $f: X \to Y$ is *light* if $f^{-1}(y)$ is finite (or totally disconnected; for surfaces X, Y it comes to the same thing) for each $y \in Y$ and *open* provided $f(u)$ is open for each open set $U \subset X$. In these terms there is the classical (1929) result as follows.

STÖILOW'S THEOREM [7, p. 103]. *If $f: M^2 \to \mathbf{C}$ is a light open map from the surface $M(\partial M = \varnothing)$ to the complex plane then there is an analytic function $\phi: R \to \mathbf{C}$, R a Riemann surface and a homeomorphism $h: R \to M^2$ such that $\phi = f \circ h$.*

We next outline the proof of Theorem C. We construct a light open map $f: \mathbf{R}^2 \to \mathbf{C}$ such that

(a) f is zero only at the sinks $v \in G$ and has degree m at v, m the weight of the sink v.

(b) The degree of f at a saddle θ is $k = k(\theta)$ where θ is a k-prong in G.

(c) f has no points of degree >1 except as in (a) and (b).

(d) The f-image of each directed edge is a (straight) ray pointed toward the origin in $\mathbf{C}$.

(e) f is proper.

Now applying Stöilow's theorem we obtain an analytic map $\phi: R \to \mathbf{C}$, and a homeomorphism $h: R \to \mathbf{R}^2$. Now R must be $\mathbf{C}$ or the interior of a disk in $\mathbf{C}$. But the latter case is ruled out as ϕ is proper. Then ϕ is a polynomial since it is entire and has poles only at ∞. But G_ϕ is just $h^{-1}(G)$ by our construction. That is, the zeros occur only at the sinks of $h^{-1}(G)$ and the other singular points are $\phi(h^{-1}(\theta))$, θ a saddle of G. Finally the solution curves of V_ϕ being those curves which map onto rays of $\mathbf{C}$ pointing toward the origin, include the directed edges of $h^{-1}(G)$.

This completes the proof of Theorem C; it remains only to construct the light open map f satisfying (a)-(e).

Construction of f. The construction will use induction on the number of saddle points in G.

Suppose there are no saddle points in the graph G. Then G consists of one root of weight m. In this case $G_f = G$ for $f(z) = z^m$.

We make the induction hypothesis that such f exist is true for dynamic graphs with less than or equal to n saddle points.

An equivalent form of the induction hypothesis which will be convenient is the following. Let G be a dynamic graph with less than or equal to n saddle points. For any disc D containing G there exists a light open map f such that (1) $f(D) = D_1 = \{z \mid |z| \leqq 1\}$, (2) f is a covering map from the boundary of D onto the boundary of D_1 and (3) f satisfies (a), (b), (c), (d) relative to G.

In order to see that the second form of the induction hypothesis follows from the first, observe that a large enough disc $\{z \mid |z| \leqq R\}$ in the range of f has for preimage a disc containing D. Adjusting f by an isotropy in the domain and by a radial isotopy in the range gives $f: D \to D_1$ with the desired G_f.

Suppose we have a dynamic graph G which is connected and acyclic, with $n+1$ saddle points. The points of G are partially ordered by the directed edges. Choose a saddle point θ at which no edge terminates. Let $\{\gamma_i\}$, $1 \leqq i \leqq k$ be the edges emanating from θ. For each γ_i, $G - \{\text{interior of } \gamma_i\}$ is the union of two dynamic graphs. Let G_i be the component of $G - \{\text{interior of } \gamma_i\}$ not containing θ. Let $\{D_i\}$, $1 \leqq i \leqq k$, be a family of pairwise disjoint discs such that D_i contains G_i, $\partial D_i \cap \gamma_i$ is a single point p_i in the interior of γ_i, and $D_i \cap \gamma_j = \phi$ for $i \neq j$. Such discs D_i can be found by taking small neighborhoods of the G_i. Since G_i is connected and acyclic, by the induction hypothesis there is a light open map $f_i: D_i \to D_1$ such that $G_i = G_{f_i}$.

Denote by q_i the terminal end of γ_i.

LEMMA. *We can alter γ_i by an isotopy so that $f(q_i) = \lambda_i \cdot f(p_i)$, for some real number $\lambda_i \geqq 0$, and so that the part of γ_i from p_i to q_i is mapped by f_i onto the radial segment from $f_i(p_i)$ to $f_i(q_i)$. Furthermore, we can assume that $f_i(p_i) = 1$ for $1 \leqq i \leqq k$.*

Proof of Lemma. Let $\{\theta_j^i\}$ be the saddle points of f_i. Then $f_i(G_i)$ is the star formed by the union of the radial segments from $f(\theta_j^i)$ to 0. We consider separately two cases: (1) q_i is a saddle of G_i and (2) q_i is a sink of G_i. In case 1, there is a unique radial segment J from a point of the circle $\{z \mid |z| = 1\}$ to $f(q_i)$. Let I be the union of all curves in $f_i^{-1}(J)$ that terminate at the saddle point q_i. There is one and only one such curve between each pair of successive edges emanating from q_i. If J contains some $f(\theta_j^i)$ then one of the curves of I will contain an edge of G_i which terminates at q_i. Suppose γ_i arrives at q_i between the two successive outward edges e_0, e_1. By property 3 of the definition of a saddle connection for the dynamic graph G, no edge of G_i can arrive at q_i between e_0 and e_1. Let $I_{0,1}$ be the curve in I that arrives between e_0 and e_1. Then $I_{0,1}$ does not contain an edge of G_i. Since G_i is connected and acyclic γ_i can be isotoped (that is, G can be isotoped, fixing $G - \gamma_i$) so that γ_i agrees with $I_{0,1}$ in D_i.

In case 2, γ_i arrives at the sink q_i and $f_i(\gamma_i \cap D_i)$ is a topological line segment proceeding from 0 to some point $p_i' \in \partial D$. Of course the map f_i, which we know exists, is only weakly related to the arc γ_i since γ_i is *not* a part of the graph G_i. We construct a more appropriate arc as follows. Choose a point p_i' near q_i in the correct "angle" or cone at q_i. Then $f(p_i') \neq 0$; let J be the line interval from 0 to $f(p_i')$. Then there is a unique lifting of J to an arc I_γ joining p_i' to q_i. This arc lifting property is essentially trivial to understand since we know f_i is a branched covering.

Let R_i be the rotation of $\mathbf{C}$ centered at $z = 0$ such that $R_i(f(p_i)) = 1$. Define $\bar{f}_i = R_i \circ G_i$. Note that $G_{\bar{f}_i} = G_f$ because R_i preserves radial lines. This completes the proof of the lemma.

We now have the situation pictured in Fig. 5.1, which we can think of as a map f defined on the union of the disks $D_1, \cdots, D_i$. Thus it is a simple matter to define a star-shaped disk D_δ^* bounded by the dotted lines in Fig. 5.1 and small arcs on the disks $D_1, \cdots, D_i$. This disk D^* is in turn mapped into the disk D' bounded by the dotted line $A = F(A_i)$ and a small arc of D, by a covering map, branched at $F(\theta)$ (see Fig. 5.2). The resulting map F is a light open map having the appropriate properties except that it is defined only on a compact disk $D_0 = D^* \cup D_1 \cup \cdots \cup D_i$. But since $F(D_0) = D \cup D'$ is a covering on the boundary, it is a simple matter to extend it to the whole plane by a covering map. This completes the induction and thus the proof of Theorem C.

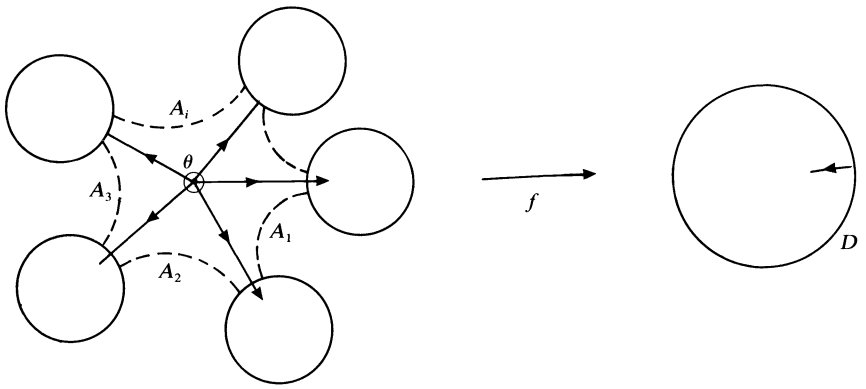


FIG. 5.1

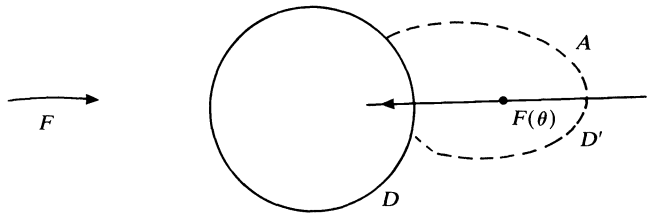


FIG. 5.2

Acknowledgements. We would like to thank Instituto Matematica Pura e Aplicada, Rio de Janeiro, Brazil, for its hospitality. It is also a pleasure to thank the excellent referee for helpful suggestions and criticism that led us to understand an error in an earlier version of our proof.

REFERENCES

[1] M. W. HIRSCH AND S. SMALE, *On algorithms for solving* $f(x) = 0$, Comm. Pure Appl. Math., 32 (1979), pp. 281-312.

[2] H. T. H. JONGEN, P. JONKER AND F. TWILT, *The continuous desingularized Newton's method for meromorphic functions*, preprint, Department of Applied Math., Twente University of Technology, the Netherlands.

[3] S. SMALE, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. N.S., 4, (1981) pp. 1-36.

[4] ———, *On the efficiency of algorithms of analysis*, Bull. Amer. Math. Soc. N.S., 13 (1985), pp. 87-122.

[5] S. STÖILOW, *Sur une théorèm topologique*, Fund. Math., 13 (1929), pp. 186-194.

[6] R. THOM, *L'équivalence d'une fonction différentiable et d'un polynome*, Topology (1964), pp. 297-304.

[7] G. T. WHYBURN, *Analytic Topology*, Princeton University Press, Princeton, NJ, 1964.

[8] PALIS-DE MELO, *Geometric Theory of Dynamical Systems*, Springer, Berlin, New York, Heidelberg, 1982.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

A MODEL EQUATION FOR VISCOELASTICITY WITH A STRONGLY SINGULAR KERNEL*

WILLIAM J. HRUSA† AND MICHAEL RENARDY‡

Abstract. In much of the mathematical work on nonlinear viscoelasticity it is assumed that the kernel (or memory function) is smooth on $[0, \infty)$. There are, however, theoretical and experimental indications that certain viscoelastic materials may be described by equations involving kernels that are singular at zero. In this paper we establish local (in time) existence of smooth solutions to a nonlinear integrodifferential equation with a singular kernel. This equation provides a model for the motion of a certain class of viscoelastic materials. Our analysis is based on energy estimates and properties of positive definite kernels.

Key words. singular kernels, viscoelasticity, local existence, energy estimates

AMS(MOS) subject classifications. 35Q99, 45K05, 73F15

1. Introduction and statement of results. Over the last decade, a significant amount of effort has been devoted to the study of nonlinear integrodifferential equations that model motions of viscoelastic materials. Most of the results obtained so far concern equations with kernels that are smooth on $[0, \infty)$. There are, however, theoretical and experimental indications that certain viscoelastic materials may be described by equations involving kernels that are singular at zero. (See, e.g., [5], [16], [18], [27], [31].) A number of interesting questions are directly linked to behavior of the kernel near zero.

Recent work on linear equations with constant coefficients ([4], [10], [14], [23], [24]) shows that singular kernels lead to smoothing of solutions. One therefore expects that nonlinear equations with singular kernels should have "nicer" existence properties than those with regular kernels. However, singular kernels lead to significant technical complications, and even questions of local existence become very delicate.

In this paper, we study the model problem

$$(1.1)_1 \qquad u_{tt}(x,t) = \chi(u_x(x,t))_x + \int_0^t a(t-\tau)\psi(u_x(x,\tau))_{x\tau}\, d\tau + f(x,t), \qquad x \in [0,1], \quad t \geqq 0,$$

$$(1.1)_2 \qquad u(0,t) = u(1,t) = 0, \qquad t \geqq 0,$$

$$(1.1)_3 \qquad u(x,0) = u_0(x), \quad u_t(x,0) = u_1(x), \quad x \in [0,1].$$

Here χ, $\psi : \mathbb{R} \to \mathbb{R}$ are assigned smooth functions, $a : (0, \infty) \to \mathbb{R}$ is a given kernel, f is a known forcing function, and u_0, u_1 are prescribed initial data. The unknown function u represents displacement. On physical grounds, it is natural to assume that a is positive, decreasing and convex, and that $\psi' > 0$. If the material in question is a fluid then $\chi' \equiv 0$, while for a solid, χ' generally is positive near equilibrium, but may change signs globally.

For physical problems arising in viscoelasticity, the integral in $(1.1)_1$ would extend from $-\infty$ to t and the history of u prior to time $t = 0$ would also be prescribed. This

* Received by the editors September 4, 1987; accepted for publication October 1, 1987. This research was sponsored by the National Science Foundation under grant DMS-8796241 and the United States Air Force under grant AFOSR-85-0307.

† Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

type of problem can be put in the form (1.1) by incorporating the part of the integral from $-\infty$ to 0 into the forcing term f. However, if the kernel a is singular then in order to ensure that the original history value problem is equivalent to the initial value problem (1.1) we must require that the limit as $t\uparrow 0$ of the given history is equal to u_0.

The behavior of smooth solutions of (1.1) (and of similar problems) is well understood when a is smooth on $[0, \infty)$, i.e., a, $a' \in AC[0, \infty)$. The situation can be described roughly as follows: if $\chi' + a(0)\psi' > 0$ and the data (u_0, u_1, f) are sufficiently regular, then (1.1) has a unique classical solution on a maximal time interval $[0, T_0)$. Under some additional assumptions (which are physically motivated) the solution of (1.1) exists globally in time, provided that the data are suitably small. On the other hand, if the data are too large then the solution will develop singularities in finite time. See, e.g., [2], [3], [9], [11], [12], [25], [28], [30], as well as the recent monograph [26] and the references cited therein.

The local existence result mentioned above can be established by a relatively simple iteration procedure that requires $\chi' + a(0)\psi' > 0$, but is otherwise insensitive to sign conditions on a and ψ. As explained in [15], [25], this procedure cannot work unless $a'(0^+)$ is finite.

The main effects of a kernel which is smooth on $[0, \infty)$ are exemplified by the special case

$$a(t) = \mu e^{-\lambda t} \tag{1.2}$$

with μ, $\lambda > 0$. When a is given by (1.2), the integrodifferential equation $(1.1)_1$ can be converted to the partial differential equation

$$u_{ttt} + \lambda u_{tt} = (\chi(u_x) + \mu\psi(u_x))_{xt} + \lambda\chi(u_x)_x + f_t + \lambda f, \tag{1.3}$$

which is studied in [7], [13]. We note that if $\chi' + \mu\phi' > 0$ then (1.3) is hyperbolic.

If we formally take a to be the Dirac delta function, then $(1.1)_1$ becomes

$$u_{tt} = \chi(u_x)_x + (\psi'(u_x)u_{xt})_x + f. \tag{1.4}$$

If $\psi' > 0$ then the initial-boundary value problem (1.4), $(1.1)_2$, $(1.1)_3$ is well posed (locally in time)—irrespective of the sign and size of χ'. Moreover, under reasonable assumptions on χ and ψ, (1.4), $(1.1)_2$, $(1.1)_3$ has a globally defined classical solution, even if the data are large ([1], [8], [17], [20]). (There are large-data global existence results for (1.4) that permit χ' to be negative. However, positivity of χ' is needed to ensure asymptotic stability.)

Although (1.4) is parabolic, solutions generally do not have more spatial regularity than the data, due to the presence of stationary singularities. Indeed, in the special case when $\chi \equiv 0$ and $\psi(\xi) = \xi$ the solution of (1.4), $(1.1)_2$, $(1.2)_3$ is given by

$$u(x, t) = u_0(x) + \int_0^t v(x, s)\, ds, \tag{1.5}$$

where v satisfies the heat equation and hence is analytic. The spatial regularity of the function u in (1.5) is precisely the same as that of u_0. In contrast with the situation concerning shocks and nonlinear hyperbolic equations, stationary singularities in solutions of (1.4) do not form on their own. In other words, such singularities originate solely from singularities in the data.

If the kernel is a function, but has a singularity at zero, we except $(1.1)_1$ to behave in an intermediate fashion, somewhere between a damped hyperbolic equation (such as (1.3)) and the parabolic equation (1.4). Some results of this nature have been established for linear problems, but relatively little is known about nonlinear equations with singular kernels.

In an earlier paper [15], we studied a problem quite similar to (1.1) under hypotheses which permit $a'(0^+) = -\infty$, but require $a(0^+)$ to be finite. (The problem considered in [15] is not exactly of the form (1.1). However, under the assumptions made in [15], the problem studied there is essentially equivalent to (1.1).) We establish here a local existence theorem for (1.1) that permits the kernel a to have an integrable singularity at zero. On the other hand, we also impose stronger monotonicity assumptions on a than those needed in [15].

Our proof employs the same basic strategy as in [15], i.e., a contraction argument based on energy estimates. However, a different function space is used and an inequality of Staffans [29] is exploited to obtain a modified chain of energy estimates.

A global existence theorem (which allows $a'(0^+) = -\infty$, but requires $a(0^+)$ to be finite) is also proved in [15] for bounded intervals and extended to unbounded intervals in [12]. This result requires the data to be small, and it is not known whether solutions develop singularities in finite time if $a'(0^+) = -\infty$ and the data are large. The problem of global existence when $a(0^+) = +\infty$ is currently under investigation (see "Note added in proof").

The only other existence results for nonlinear problems with singular kernels that we know of are the works of Londen [19] and Engler [6]. Londen establishes the existence of weak solutions to an abstract integrodifferential equation. His existence theorem can be applied to (1.1) in the special case when χ is a scalar multiple of ψ. Engler establishes the existence of weak solutions to the equations of motion for a class of viscoelastic fluids.

Concerning the kernel a we require

(a)
$$a \in L^1(0, \infty),$$
$$a \geqq 0, \qquad a' \leqq 0, \qquad a'' \geqq 0, \qquad a''' \leqq 0,$$

and we make the following assumptions of smoothness:

(s1) $$\chi, \psi \in C^3(\mathbb{R});$$

(s2) $$u_0 \in H^3(0, 1), \qquad u_1 \in H^2(0, 1);$$

(s3) $$f, f_x, f_{xx} \in L^2_{\text{loc}}([0, \infty); L^2(0, 1)).$$

In order to obtain a smooth solution of (1.1) we need the data (u_0, u_1 and f) to be compatible with the boundary conditions; for technical reasons we make the following rather strong compatibility assumption:

(c)
$$u_0(0) = u_0(1) = u_1(0) = u_1(1) = 0,$$
$$u_0''(0) = u_0''(1) = 0,$$
$$f(0, t) = f(1, t) = 0 \quad \text{a.e. } t \geqq 0,$$

which guarantees that the data admit smooth, spatially periodic, odd extensions (of period 2). Finally, in order to ensure the evolutionarity of equation (1.1) we require

(e) $$\psi'(\xi) > 0, \quad \chi'(\xi) + a(0^+)\psi'(\xi) > 0 \quad \forall \xi \in \mathbb{R}.$$

Observe that if $a(0^+) = +\infty$ then (e) imposes no restrictions on χ'. In this regard, when $a(0^+) = +\infty$ the memory term in $(1.1)_1$ has the same effect as the Newtonian viscosity $(\psi'(u_x)u_{xt})_x$.

The hypotheses of primary interest are (a) and (e). The inequalities in (a) are assumed to hold in the sense of distributions. Of course this guarantees that the kernel a has a certain amount of smoothness on $(0, \infty)$ in the classical sense. More precisely, it follows from (a) that $a \in C^1(0, \infty)$ and that a' is locally Lipschitz continuous on

$(0, \infty)$. The main difference between (a) and the corresponding assumption in [15] can be described roughly as follows: in [15] it is also required that $a' \in L^1(0, \infty)$ (and hence that $a(0^+)$ is finite), but no condition on a''' is needed.

Our main result is the following theorem.

THEOREM. *Assume that* (a), (s1)-(s3), (c) *and* (e) *hold. Then the initial-boundary value problem* (1.1) *has a unique solution u on a maximal time interval* $[0, T_0)$, $T_0 > 0$, *with*

$$u, u_x, u_t, u_{xx}, u_{xt}, u_{xxx}, u_{xxt} \in L^\infty_{\mathrm{loc}}([0, T_0); L^2(0, 1)), \tag{1.6}_1$$

$$u_{tt}, u_{xtt} \in L^2_{\mathrm{loc}}([0, T_0); L^2(0, 1)). \tag{1.6}_2$$

Moreover, if

$$\operatorname*{ess\text{-}sup}_{t \in [0, T_0)} \int_0^1 \{u_{xxx}^2 + u_{xxt}^2\}(x, t)\, dx + \int_0^{T_0} \int_0^1 u_{xtt}^2(x, t)\, dx\, dt < \infty, \tag{1.7}$$

then $T_0 = \infty$.

Remarks. (1) It follows from (1.6) and standard embedding theorems that u, u_x, u_t, u_{xx} and u_{xt} are continuous on $[0, 1] \times [0, T_0)$. Moreover, since f vanishes at the endpoints, one can show that the solution satisfies $u_{xx}(0, t) = u_{xx}(1, t) = 0$ for all $t \in [0, T_0)$.

(2) A similar existence theorem holds for the pure initial value problem on all of space.

(3) The question of optimal regularity of the solution of (1.1) appears to be rather delicate. One expects that a singularity in a will lead to smoothing in the temporal direction. However, one should not expect spatial smoothing because equation $(1.1)_1$ permits stationary singularities.

The paper is organized as follows. In § 2 we discuss some preliminary material concerning the kernel. Then in § 3, we establish an existence theorem for a linear integrodifferential equation with variable coefficients. Finally, in § 4, we use the results of § 3 to prove the theorem stated above. In §§ 3 and 4 we emphasize those features of our proofs that differ from [15]; details of arguments that are very similar to ones in [15] will be omitted.

2. Preliminaries. This section contains some preliminary material that is needed for the proof of our theorem. Let X be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. For each $b \in L^1_{\mathrm{loc}}[0, \infty)$, $T > 0$, and $\Phi \in L^2([0, T]; X)$ we set

$$Q(\Phi, t, b) := \int_0^t \left\langle \Phi(s), \int_0^s b(s - \tau)\Phi(\tau)\, d\tau \right\rangle ds \quad \forall t \in [0, T]. \tag{2.1}$$

We denote by $H^1([0, T]; X)$ the set of all $\Phi \in L^2([0, T]; X)$ such that $\Phi' \in L^2([0, T]; X)$, where Φ' is the (distributional) derivative of Φ.

Our energy estimates make crucial use of several properties of positive definite kernels. For the sake of completeness we recall a few basic concepts. A real-valued function $b \in L^1_{\mathrm{loc}}[0, \infty)$ is said to be positive definite (or of positive type) if

$$\int_0^t v(s) \int_0^s b(s - \tau)v(\tau)\, d\tau\, ds \geqq 0 \qquad \forall t \geqq 0 \tag{2.2}$$

for every $v \in C[0, \infty)$; b is called strongly positive definite if there is a $\lambda > 0$ such that the function $t \to b(t) - \lambda e^{-t}$ is positive definite. As the terminology suggests, every strongly positive definite kernel is positive definite.

The definition of a positive definite function is not easy to check directly. We quote a well-known sufficient condition. If $b \in L^1_{\mathrm{loc}}[0, \infty)$ satisfies

$$b \geqq 0, \qquad b' \leqq 0, \qquad b'' \geqq 0, \tag{2.3}$$

then b is positive definite; if, in addition, the measure b'' has a nontrivial absolutely continuous component, then b is strongly positive definite. We note that if b is positive definite and $\Phi \in L^2([0, T]; X)$ then

$$Q(\Phi, t, b) \geqq 0 \quad \forall t \in [0, T]. \tag{2.4}$$

See, for example, [22], [29] for more information on these matters.

In our existence proof for the linearized problem we shall employ "shifted" kernels. For each $\delta > 0$ we defined $a_\delta : [0, \infty) \to \mathbb{R}$ by

$$a_\delta(s) := a(s + \delta) \quad \forall s \geqq 0. \tag{2.5}$$

It follows from (a) and (2.5) that

$$a_\delta \geqq 0, \qquad a_\delta' \leqq 0, \qquad a_\delta'' \geqq 0, \qquad a_\delta''' \leqq 0, \tag{2.6}$$

$$a_\delta, a_\delta', a_\delta'' \in L^1(0, \infty) \tag{2.7}$$

and

$$\|a_\delta\|_{L^1} \leqq \|a\|_{L^1} \tag{2.8}$$

for every $\delta > 0$. Observe that $a_\delta \to a$ pointwise (and in L^1) as $\delta \downarrow 0$. The use of shifted kernels is not essential, but it provides a simple way of constructing approximate problems which are known to have solutions.

In view of (2.8), the following estimate is an immediate consequence of an inequality of Staffans. (See Lemma 1 and Theorem 2(iii) of [29].)

LEMMA 1. *Assume that* (a) *holds and let* $A := 5\|a\|_{L^1}$. *Then, for every* δ, $T > 0$, *and* $\Phi \in L^2([0, T]; X)$ *we have*

$$\int_0^t \left\| \int_0^s a_\delta(s - \tau)\Phi(\tau)\, d\tau \right\|^2 ds \leqq A \cdot Q(\Phi, t, a_\delta) \quad \forall t \in [0, T]. \tag{2.9}$$

Lemma 1 of [29] is formulated for the case when Φ is a continuous scalar-valued function. However, the same proof applies under the present circumstances.

The next result provides a useful lower bound for $Q(\Phi', t, a_\delta)$.

LEMMA 2. *Assume that* (a) *holds and let* δ_0, $\varepsilon > 0$ *be given. Then there is a constant* $C = C(\delta_0, \varepsilon)$ *such that for every* $T > 0$, $\delta \in (0, \delta_0]$, *and* $\Phi \in H^1([0, T]; X)$ *we have*

$$\begin{aligned} Q(\Phi', t, a_\delta) \geqq{}& \frac{1}{2}(a(\delta_0) - \varepsilon)\|\Phi(t) - \Phi(0)\|^2 \\ & - C \int_0^t \|\Phi(s) - \Phi(0)\|^2\, ds \quad \text{a.e. } t \in [0, T]. \end{aligned} \tag{2.10}$$

Proof. It is clear that the conclusion of the lemma holds if $a(\delta_0) = 0$, so we assume that $a(\delta_0) > 0$. We note that a_{δ_0} is then strongly positive definite and that $a_{\delta_0}, a_{\delta_0}' \in L^1(0, \infty)$.

By virtue of Lemma 2.3 of [15], there is a constant $C = C(\delta_0, \varepsilon)$ such that for every $T > 0$ and every $\Phi \in H^1([0, T]; X)$ we have

$$\begin{aligned} Q(\Phi', t, a_{\delta_0}) \geqq{}& \left(\frac{1}{2}a(\delta_0) - \varepsilon\right)\|\Phi(t) - \Phi(0)\|^2 \\ & - C \int_0^t \|\Phi(s) - \Phi(0)\|^2\, ds \quad \forall t \in [0, T]. \end{aligned} \tag{2.11}$$

(To apply the results of [15], we set $\Phi(t)=\Phi(0)$ for $t<0$, put $u(t)=\Phi(t)-\Phi(0)$, use a_{δ_0} in place of a, and let $h\downarrow 0$.) It follows from (a) and (2.5) that $a_\delta - a_{\delta_0}$ is positive definite for all $\delta\in(0,\delta_0]$ and consequently

$$Q(\Phi', t, a_\delta)\geqq Q(\Phi', t, a_{\delta_0}) \quad \forall \delta\in(0,\delta_0]. \tag{2.12}$$

This completes the proof.

3. Linear problem. The proof of our theorem is based on an iteration scheme which involves linear problems of the form

$$u_{tt}(x,t)=\gamma(x,t)u_{xx}(x,t)+\int_0^t a(t-\tau)[\beta(x,\tau)u_{xx}(x,\tau)]_\tau\,d\tau+f(x,t) \tag{3.1}_1$$
$$\forall x\in[0,1],\ t\in[0,T],$$

$$u(0,t)=u(1,t)=0, \tag{3.1}_2$$

$$u(x,0)=u_0(x),\ u_t(x,0)=u_1(x). \tag{3.1}_3$$

In this section we establish existence and an a priori estimate for (3.1). Concerning the coefficients γ and β we assume

$$\gamma,\gamma_x,\gamma_t,\gamma_{xx},\gamma_{xt},\beta,\beta_x,\beta_t,\beta_{xx},\beta_{xt}\in L^\infty([0,T];L^2(0,1)), \quad \gamma_{tt},\beta_{tt}\in L^2([0,T];L^2(0,1)); \tag{s1*}$$

$$\beta(x,t)\geqq\underline{\beta}>0,\ \gamma(x,t)+a(0^+)\beta(x,t)\geqq\underline{\lambda}>0 \quad \forall x\in[0,1],\ t\in[0,T]; \tag{e*}$$

$$\gamma_x(0,t)=\gamma_x(1,t)=\beta_x(0,t)=\beta_x(1,t)=0 \quad \forall t\in[0,T]. \tag{c*}$$

Our assumptions on the kernel and the data are the same as for the nonlinear problem.

For the purpose of stating an a priori estimate we define

$$U_0:=\int_0^1\{u_0'''(x)^2+u_1''(x)^2\}\,dx, \tag{3.2}$$

$$F:=\int_0^T\int_0^1 f_{xx}^2(x,t)\,dx\,dt, \tag{3.3}$$

$$B:=\int_0^1\{\gamma^2+\gamma_x^2+\beta^2+\beta_x^2\}(x,0)\,dx, \tag{3.4}$$

$$\Gamma:=\operatorname*{ess\text{-}sup}_{t\in[0,T]}\int_0^1\{\gamma^2+\gamma_x^2+\gamma_t^2+\gamma_{xx}^2+\gamma_{xt}^2+\beta^2+\beta_x^2+\beta_t^2+\beta_{xx}^2+\beta_{xt}^2\}(x,t)\,dx. \tag{3.5}$$

LEMMA 3. *Let $T>0$ be given and assume that* (a), (s1*), (s2), (s3), (c), (c*) *and* (e*) *hold. Then, the initial boundary value problem* (3.1) *has a unique solution u with*

$$u,u_x,u_t,u_{xx},u_{xt},u_{xxx},u_{xxt}\in L^\infty([0,T];L^2(0,1)), \tag{3.6}_1$$

$$u_{tt},u_{xtt}\in L^2([0,T];L^2(0,1)). \tag{3.6}_2$$

Moreover, this solution obeys the a priori estimate

$$\operatorname*{ess\text{-}sup}_{t\in[0,T]}\int_0^1\{u_{xxx}^2+u_{xxt}^2\}(x,t)+\int_0^T\int_0^1 u_{xtt}^2(x,t)\,dx\,dt \leqq K\{F+(1+B+BT)U_0\}\exp[K(1+\Gamma)T], \tag{3.7}$$

where K is a constant that depends on $\underline{\beta}$ and $\underline{\lambda}$, but is independent of U_0, F, B, Γ and T.

Proof. We use the shifted kernels a_δ (which are smooth on $[0,\infty)$) in place of a to construct approximate solutions $u^{(\delta)}$. We shall also approximate f by functions with more temporal regularity so that standard theory of equations with smooth kernels can be used to solve for the $u^{(\delta)}$. Let us set

$$f(x,t)=f(x,-t) \quad \text{for } t<0 \tag{3.8}$$

and

$$f^{(\delta)}(x,t):=\int_{-\infty}^{\infty} J_\delta(t-\tau)f(x,\tau)\,d\tau \tag{3.9}$$

where J_δ is a standard mollifier on $\mathbb{R}$.

We replace (3.1) with

$$u_{tt}^{(\delta)}(x,t)=\gamma(x,t)u_{xx}^{(\delta)}(x,t)+\int_0^t a_\delta(t-\tau)[\beta u_{xx}^{(\delta)}]_t(x,\tau)\,d\tau+f^{(\delta)}(x,t), \tag{3.10}$$

$$x\in[0,1], \quad t\in[0,T],$$

$$u^{(\delta)}(0,t)=u^{(\delta)}(1,t)=0, \tag{3.10}$$

$$u^{(\delta)}(x,0)=u_0(x), \qquad u_t^{(\delta)}(x,0)=u_1(x), \tag{3.10}$$

and we choose $\delta_0>0$ and small enough so that

$$\underline{\alpha}:=\inf_{\substack{x\in[0,1]\\ t\in[0,T]}}[\gamma(x,t)+a(\delta_0)\beta(x,t)]>0. \tag{3.11}$$

Observe that

$$\gamma(x,t)+a_\delta(0)\beta(x,t)\geqq\underline{\alpha} \quad \forall x\in[0,1], \quad t\in[0,T], \quad \delta\in(0,\delta_0]. \tag{3.12}$$

An integration by parts in (3.10)$_1$ produces

$$\begin{aligned} u_{tt}^{(\delta)}(x,t)&=[\gamma+a_\delta(0)\beta]u_{xx}^{(\delta)}(x,t)+\int_0^t a_\delta'(t-\tau)[\beta u_{xx}^{(\delta)}](x,\tau)\,d\tau \\ &=f^{(\delta)}(x,t)+a_\delta(t)\beta(x,0)u_0''(x). \end{aligned} \tag{3.13}$$

It follows from a standard argument (cf., e.g., [26]) that for each $\delta\in(0,\delta_0]$, the initial-boundary value problem (3.13), (3.10)$_2$, (3.10)$_3$ has a unique solution $u^{(\delta)}$ with

$$u^{(\delta)}, u_x^{(\delta)}, u_t^{(\delta)}, u_{xx}^{(\delta)}, u_{xt}^{(\delta)}, u_{tt}^{(\delta)}, u_{xxx}^{(\delta)}, u_{xxt}^{(\delta)}, u_{xtt}^{(\delta)}, u_{ttt}^{(\delta)}\in C([0,T];L^2(0,1)). \tag{3.14}$$

Moreover, this solution satisfies

$$u_{xx}^{(\delta)}(0,t)=u_{xx}^{(\delta)}(1,t)=0 \quad \forall t\in[0,T]. \tag{3.15}$$

It is clear that $u^{(\delta)}$ is also a solution of (3.10). Our objective is to obtain a priori bounds for $u^{(\delta)}$ that guarantee the existence of a sequence $\{\delta_n\}_{n=1}^\infty$ tending to zero such that $u^{(\delta_n)}$ converges to a solution of (3.1). For the purpose of obtaining these bounds, we shall extend $u^{(\delta)}$, u_0, u_1, $f^{(\delta)}$, γ and β periodically in space, in such a way that the extended functions are smooth and (3.10)$_1$, (3.10)$_3$ hold for all $x\in\mathbb{R}$. This will allow us to employ spatial difference operators.

We extend $u^{(\delta)}$, u_0, u_1 and $f^{(\delta)}$ (spatially) to be odd periodic functions of period 2, and we extend γ and β to be even periodic functions of period 2. By virtue of (c) and (c*), the extended functions have the same regularity as the original functions. Moreover, (3.10)$_1$ and (3.10)$_3$ are satisfied for all $x\in\mathbb{R}$.

In the estimates that follow we use K to denote a generic positive constant that can be chosen independently of U_0, F, B, Γ, T and δ. Moreover, to simplify the notation we suppress the superscript δ on u and f. We note that the elementary inequality

$$|CD| \leqq \eta C^2 + \frac{1}{4\eta} D^2 \quad \forall C, D \in \mathbb{R}, \eta > 0 \tag{3.16}$$

will be exploited in several places.

For each $h > 0$, we define the spatial difference operator D_h by

$$(D_h v)(x, t) := v(x+h, t) - v(x, t). \tag{3.17}$$

Applying D_h to $(3.10)_1$ we obtain

$$D_h u_{tt} = D_h[\gamma u_{xx}] + \int_0^t a_\delta(t-\tau) D_h[(\beta u_{xx})_t](x, \tau)\, d\tau + D_h f. \tag{3.18}$$

We multiply (3.18) by $D_h[(\beta u_{xx})_t]$ and integrate over space and time. After several integrations by parts we let $h \downarrow 0$ and obtain

$$\begin{aligned} &\frac{1}{2}\int_0^1 \{\beta u_{xxt}^2 + \beta\gamma u_{xxx}^2\}(x, t)\, dx + \lim_{h\downarrow 0} \frac{1}{h^2} Q(D_h[(\beta u_{xx})_t], t, a_\delta) \\ &\qquad = \int_0^t \int_0^1 (\beta_t u_{xx})_x u_{xtt}(x, s)\, dx\, ds + R_1(t) \quad \forall t \in [0, T], \end{aligned} \tag{3.19}$$

where Q is given by (2.1) with $X := L^2(0, 1)$ and R_1 is given by

$$\begin{aligned} R_1(t) = &\frac{1}{2}\int_0^1 \{\beta u_{xxt}^2 + \beta\gamma u_{xxx}^2\}(x, 0)\, dx \\ &+ \int_0^t \int_0^1 \left\{\frac{1}{2}\beta_t u_{xxt}^2 + \frac{1}{2}\beta\gamma_t u_{xxx}^2 - \frac{1}{2}\beta_t \gamma u_{xxx}^2 \right. \\ &\quad - \beta_x \gamma u_{xxx} u_{xxt} + \beta\gamma_x u_{xxx} u_{xxt} - \beta_t \gamma_x u_{xx} u_{xxx} + \beta\gamma_{xx} u_{xx} u_{xxt} \\ &\quad \left. - \beta_{xt}\gamma u_{xx} u_{xxx} - \beta_{xt}\gamma_x u_{xx}^2 + \beta f_{xx} u_{xxt} + \beta_t f_{xx} u_{xx} \right\}(x, s)\, dx\, ds. \end{aligned} \tag{3.20}$$

It is not evident a priori that $\lim_{h\downarrow 0}(1/h^2)\ Q(D_h[(\beta u_{xx})_t], t, a_\delta)$ exists for a.e. $t \in [0, T]$. However, the limits of each of the other terms involved in the derivation of (3.19) exist for a.e. $t \in [0, T]$. Consequently, the limit in question exists for a.e. $t \in [0, T]$.

We choose ε sufficiently small and apply Lemma 2 (and some straightforward calculations) to conclude that the left side of (3.19) is bounded below by

$$\int_0^1 \left\{\frac{1}{2}\underline{\beta} u_{xxt}^2 + \frac{1}{4}\underline{\alpha\beta} u_{xxx}^2\right\}(x, t)\, dx - K \cdot R_2(t) \tag{3.21}$$

where

$$\begin{aligned} R_2(t) := &\int_0^1 \{[(\beta u_{xx})_x(x, 0)]^2 + \beta_x^2 u_{xx}^2(x, t)\}\, dx \\ &+ \int_0^t \int_0^1 [(\beta u_{xx})_x(x, s) - (\beta u_{xx})_x(x, 0)]^2\, dx\, ds. \end{aligned} \tag{3.22}$$

It therefore follows from (3.19) that

$$\int_0^1 \{u_{xxx}^2 + u_{xxt}^2\}(x, t)\, dx \leqq K \cdot R_1(t) + K \cdot R_2(t) + K \int_0^t \int_0^1 (\beta_t u_{xx})_x u_{xtt}(x, s)\, dx\, ds. \tag{3.23}$$

We now divide (3.18) by h and employ Lemma 1 to conclude that

$$\begin{aligned} &\frac{1}{h^2} \int_0^t \int_0^1 [D_h u_{tt} - D_h(\gamma u_{xx}) - D_h f]^2(\cdot, s)\, dx\, ds \\ &\leqq \frac{1}{h^2} \int_0^t \int_0^1 \left(\int_0^s a_\delta(s-\tau) D_h[(\beta u_{xx})_t](\cdot, \tau)\, d\tau \right) dx\, ds \\ &\leqq \frac{A}{h^2} Q(D_h[(\beta u_{xx})_t], t, a_\delta). \end{aligned} \tag{3.24}$$

Letting $h \downarrow 0$ in (3.24) we obtain

$$\int_0^t \int_0^1 [u_{xtt} + (\gamma u_{xx})_x - f_x]^2(x, s)\, dx\, ds \leqq A \cdot \lim_{h \downarrow 0} \frac{1}{h^2} Q(D_h[(\beta u_{xx})_t], t, a_\delta). \tag{3.25}$$

It follows easily from (3.25) that

$$\int_0^t \int_0^1 u_{xtt}^2(x, s)\, dx\, ds \leqq K \cdot R_3(t) + K \cdot \lim_{h \downarrow 0} \frac{1}{h^2} Q(D_h[(\beta u_{xx})_t], t, a) \tag{3.26}$$

where

$$R_3(t) := \int_0^t \int_0^1 \{[(\gamma u_{xx})_x]^2 + f_x^2\}(x, s)\, dx\, ds. \tag{3.27}$$

The combination of (3.19), (3.23) and (3.25) yields the estimate

$$\begin{aligned} &\int_0^1 \{u_{xxx}^2 + u_{xxt}^2\}(x, t)\, dx + \int_0^t \int_0^1 u_{xtt}^2(x, s)\, dx\, ds \\ &\leqq K\{R_1(t) + R_2(t) + R_3(t)\} + K \int_0^t \int_0^1 (\beta_t u_{xx})_x u_{xtt}(x, s)\, dx\, ds. \end{aligned} \tag{3.28}$$

We apply (3.16) (with η sufficiently small) to the integrand on the right of (3.28) to obtain

$$\int_0^1 \{u_{xxx}^2 + u_{xxt}^2\}(x, t)\, dx + \int_0^t \int_0^1 u_{xtt}^2(x, s)\, dx\, ds \leqq K\{R_1(t) + R_2(t) + R_3(t) + R_4(t)\} \tag{3.29}$$

with

$$R_4(t) := \int_0^t \int_0^1 [(\beta_t u_{xx})_x]^2(x, s)\, dx\, ds. \tag{3.30}$$

To proceed further, we introduce

$$E[u](t) := \operatorname*{ess\text{-}sup}_{s \in [0,t]} \int_0^1 \{u_{xxx}^2 + u_{xxt}^2\}(x, s)\, ds. \tag{3.31}$$

Using calculations similar to those on pages 209–211 of [15] to bound the remainder terms $R_i(t)$, we arrive at the following estimate:

$$E[u](t)+\int_0^t\int_0^1 u_{xtt}^2(x,s)\,dx\,ds \leqq K\{F+(1+B+BT)U_0\} + K(1+\Gamma)\int_0^t E[u](s)\,ds. \tag{3.32}$$

(In our deviation of (3.32) we have exploited Poincaré's inequality; this is not essential, but it leads to a slightly simpler estimate.) Gronwall's inequality applied to (3.32) yields

$$E[u](T)+\int_0^T\int_0^1 u_{xtt}^2(x,s)\,dx\,ds \leqq K\{F+(1+B+BT)U_0\}\exp[K(1+\Gamma)T]. \tag{3.33}$$

It follows from (3.33) that $u_{xxx}^{(\delta)}$, $u_{xxt}^{(\delta)}$ are bounded in $L^\infty([0,T];L^2(0,1))$ and $u_{xtt}^{(\delta)}$ is bounded in $L^2([0,T];L^2(0,1))$ independently of δ. Therefore, there is a function $u:[0,1]\times[0,T]\to\mathbb{R}$ and sequence $\delta_n\downarrow 0$ such that

$$u^{(\delta_n)}, u_x^{(\delta_n)}, u_t^{(\delta_n)}, u_{xx}^{(\delta_n)}, u_{xt}^{(\delta_n)}, u_{xxx}^{(\delta_n)}, u_{xxt}^{(\delta_n)} \to u, u_x, \text{ etc.} \quad \text{weakly * in } L^\infty([0,T];L^2(0,1)), \tag{3.34}_1$$

$$u_{tt}^{(\delta_n)}, u_{xtt}^{(\delta_n)} \to u_{tt}, u_{xtt} \quad \text{weakly in } L^2([0,T];L^2(0,1)). \tag{3.34}_2$$

It follows from (3.34) and the convergence properties of a_δ and $f^{(\delta)}$ that u is a solution of (3.1) and that u satisfies the estimate (3.33). The uniqueness of a solution with the regularity (3.6) follows from a straightforward argument.

4. Proof of the theorem. For $M, T>0$ we denote by $Z(M,T)$ the set of all functions $w:[0,1]\times[0,T]\to\mathbb{R}$ satisfying

$$w, w_x, w_t, w_{xx}, w_{xt}, w_{xxx}, w_{xxt} \in L^\infty([0,T];L^2(0,1)), \tag{4.1}_1$$

$$w_{tt}, w_{xtt} \in L^2([0,T];L^2(0,1)), \tag{4.1}_2$$

$$w(0,t)=w(1,t)=0 \quad \forall t\in[0,T], \tag{4.1}_3$$

$$w_{xx}(0,t)=w_{xx}(1,t)=0 \quad \forall t\in[0,T], \tag{4.1}_4$$

$$w(x,0)=u_0(x) \quad \forall x\in[0,1], \tag{4.1}_5$$

$$\operatorname*{ess\text{-}sup}_{t\in[0,T]}\int_0^1\{w_{xxx}^2+w_{xxt}^2\}(x,t)\,dx+\int_0^T\int_0^1 w_{xtt}^2(x,t)\,dx\,dt \leqq M^2. \tag{4.1}_6$$

We note that $Z(M,T)$ is nonempty if M is sufficiently large.

We assume temporarily that

$$\inf_{\xi\in\mathbb{R}}\psi'(\xi)>0, \qquad \inf_{\xi\in\mathbb{R}}[\chi'(\xi)+a(0^+)\psi'(\xi)]>0. \tag{4.2}$$

(As in [15], this assumption will be removed later.) Identifying γ with $\chi'(w_x)$ and β with $\psi'(w_x)$, it follows from Lemma 3 that for $w\in Z(M,T)$, the initial value problem

$$u_{tt}(x,t)=\chi'(w_x)u_{xx}(x,t)+\int_0^t a(t-\tau)[\psi'(w_x)u_{xx}(x,\tau)]_\tau\,d\tau+f(x,t), \quad x\in[0,1], \quad t\in[0,T], \tag{4.3}_1$$

$$u(0,t)=u(1,t)=0, \quad t\in[0,T], \tag{4.3}_2$$

$$u(x,0)=u_0(x), \quad u_t(x,0)=u_1(x), \quad x\in[0,1], \tag{4.3}_3$$

has a unique solution satisfying (3.6). By virtue of (4.2), the corresponding $\underline{\beta}$ and $\underline{\lambda}$ can be chosen independently of M and T.

Let S be the mapping that carries $w \in Z(M, T)$ into the solution of (4.1). We want to show that for appropriately chosen M and T, S has a unique fixed point in $Z(M, T)$; such a fixed point is obviously a solution of (1.1) on $[0, 1]\times[0, T]$.

Existence of the desired fixed point will be established by means of the contraction mapping principle. For this purpose we equip $Z(M, T)$ with the complete metric ρ defined by

$$\rho(w, \bar{w})^2 := \max_{t\in[0,T]} \int_0^1 \{(w_{xx} - \bar{w}_{xx})^2 + (w_{xt} - \bar{w}_{xt})^2\}(x, t)\, dx + \int_0^T \int_0^1 (\bar{w}_{tt} - \bar{w}_{tt})^2(x, t)\, dx\, dt. \tag{4.4}$$

It follows from the definition of $Z(M, T)$ and the a priori estimate of Lemma 3 that S maps $Z(M, T)$ into itself if M is sufficiently large and T is sufficiently small relative to M.

Let M, $T>0$ and w, $\bar{w} \in Z(M, T)$ be given and put $u := Sw$, $\bar{u} := S\bar{w}$, $W := w - \bar{w}$, $U := u - \bar{u}$. A simple calculation shows that U satisfies

$$U_{tt} = \chi'(w_x) U_{xx} + [\chi'(w_x) - \chi'(\bar{w}_x)]\bar{u}_{xx} + \int_0^t a(t-\tau)[\psi'(w_x) U_{xx}(x, \tau) + (\psi'(w_x) - \psi'(\bar{w}_x))\bar{u}_{xx}(x, \tau)]_\tau\, d\tau, \tag{4.5}_1$$

$$U(0, t) = U(1, t) = 0, \tag{4.5}_2$$

$$U(x, 0) = U_t(x, 0) = 0. \tag{4.5}_3$$

To show that S is contractive we first multiply $(4.5)_1$ by Φ_t, where

$$\Phi := \psi'(w_x) U_{xx} + [\psi'(w_x) - \psi'(\bar{w}_x)]\bar{u}_{xx}, \tag{4.6}$$

integrate the resulting equation over space and time, and exploit Lemma 2 (as in § 3). This yields an estimate for

$$\int_0^1 \{U_{xx}^2 + U_{xt}^2\}(x, t)\, dx + Q(\Phi_t, t, a). \tag{4.7}$$

We then apply Lemma 1 to $(4.5)_1$ to obtain an estimate for

$$\int_0^T \int_0^1 U_{tt}^2(x, t)\, dx\, dt. \tag{4.8}$$

Combining the estimates obtained for the quantities in (4.7) and (4.8) and employing calculations very similar to those in § 4 of [15] we obtain an inequality of the form

$$\rho(Sw, S\bar{w}) \leqq P(M, T) \exp[R(M, T)] \tag{4.9}$$

(valid for M large and T small) where P, $R:[0, \infty)\times[0, \infty)\to[0, \infty)$ are continuous functions with $P(M, 0) = 0 \forall M > 0$. If we fix M sufficiently large and then choose T sufficiently small relative to M then S maps $Z(M, T)$ into itself, and (4.9) guarantees that S is strictly contractive with respect to ρ. The rest of the proof can be carried out as in [15], and we omit the details.

Note added in proof. A forthcoming article of M. Renardy ("Coercive estimates and existence of solutions for a model of one-dimensional viscoelasticity with a nonintegrable memory function," submitted to *J. Integral Equations Appl.*) develops

an alternative existence proof based on coercive properties of the linearized problem. Global existence for small data and local existence for large data are established for an equation of motion more general than $(1.1)_1$. Roughly speaking, the analogue of the kernel a is required to have a singularity at zero that is at least as strong as a negative power of t.

REFERENCES

[1] C. M. DAFERMOS, *The mixed initial-boundary value problem for the equations of one-dimensional nonlinear viscoelasticity*, J. Differential Equations, 6 (1969), pp. 71–86.

[2] C. M. DAFERMOS AND J. A. NOHEL, *A nonlinear hyperboolic Volterra equation in viscoelasticity*, Amer. J. Math., supplement (1981), pp. 87–116.

[3] C. M. DAFERMOS, *Development of singularities in the motion of materials with fading memory*, Arch. Rational Mech. Anal., 91 (1986), pp. 193–205.

[4] W. DESCH AND R. GRIMMER, *Smoothing properties of linear Volterra integrodifferential equations*, this Journal, submitted.

[5] M. DOI AND S. F. EDWARDS, *Dynamics of concentrated polymer systems*, J. Chem. Soc. Faraday, 74 (1978), pp. 1789–1832; 75 (1979), pp. 38–54.

[6] H. ENGLER, *Weak solutions of a class of quasilinear hyperbolic integro-differential equations describing viscoelastic materials*, Arch. Rational Mech. Anal., submitted.

[7] J. M. GREENBERG, *A priori estimates for flows in dissipative materials*, J. Math. Anal. Appl., 60 (1977), pp. 617–630.

[8] J. M. GREENBERG, R. C. MACCAMY AND V. J. MIZEL, *On the existence, uniqueness, and stability of solutions of the equation* $\sigma'(u_x)u_{xx}+\lambda u_{xxt}=\rho_0 u_{tt}$, J. Math. Mech., 17 (1968), pp. 707–728.

[9] G. GRIPENBERG, *Nonexistence of smooth solutions for shearing flows in a nonlinear viscoelastic fluid*, this Journal, 13 (1982), pp. 954–961.

[10] K. B. HANNSGEN AND R. L. WHEELER, *Behavior of the solutions of a Volterra equation as a parameter tends to infinity*, J. Integral Equations, 7 (1984), pp. 229–237.

[11] H. HATTORI, *Breakdown of smooth solutions in dissipative nonlinear hyperbolic equations*, Quart. Appl. Math., 40 (1982/83), pp. 113–127.

[12] W. J. HRUSA, *Some remarks on the Cauchy problem in one-dimensional nonlinear viscoelasticity*, preprint.

[13] W. J. HRUSA AND J. A. NOHEL, *Global existence and asymptotics in one-dimensional nonlinear viscoelasticity*, in Trends and Applications of Pure Mathematics to Mechanics, Springer Lecture Notes in Physics 195, P. G. Ciarlet and M. Roseau, eds., 1984, pp. 165–187.

[14] W. J. HRUSA AND M. RENARDY, *On wave propagation in linear viscoelasticity*, Quart. Appl. Math., 43 (1985), pp. 237–254.

[15] ———, *On a class of quasilinear partial integrodifferential equations with singular kernels*, J. Differential Equations, 64 (1986), pp. 195–220.

[16] D. D. JOSEPH, O. RICCIUS AND M. ARNEY, *Shear wave speeds and elastic moduli for different liquids*, II. *Experiments*, J. Fluid Mech., 171 (1986), pp. 309–338.

[17] YA. I. KANEL', *A model system of equations for the one-dimensional motion of a gas*, Differential Equations, 4 (1968), pp. 374–380.

[18] H. M. LAUN, *Description of the nonlinear shear behaviour of a low density polyethylene melt by means of an experimentally determined strain dependent memory function*, Rheol. Acta, 17 (1978), pp. 1–15.

[19] S.-O. LONDEN, *An existence result on a Volterra equation in a Banach space*, Trans. Amer. Math. Soc., 235 (1978), pp. 285–304.

[20] R. C. MACCAMY, *Existence, uniqueness and stability of solutions of the equation* $u_{tt}=\partial/\partial x(\sigma(u_x)+\lambda(u_x)u_{xt})$, Indiana Univ. Math. J., 20 (1970), pp. 231–238.

[21] ———, *A model for one-dimensional nonlinear viscoelasticity*, Quart. Appl. Math., 35 (1977), pp. 21–33.

[22] J. A. NOHEL AND D. F. SHEA, *Frequency domain methods for Volterra equations*, Adv. Math., 22 (1976), pp. 278–304.

[23] J. PRÜSS, *Positivity and regularity of hyperbolic Volterra equations in Banach spaces*, Math. Ann., submitted.

[24] M. RENARDY, *Some remarks on the propagation and non-propagation of discontinuities in linearly viscoelastic liquids*, Rheol. Acta, 21 (1982), 251–254.

[25] ———, *Recent developments and open problems in the mathematical theory of viscoelasticity*, in Viscoelasticity and Rheology, A. S. Lodge, M. Renardy and J. A. Nohel, eds., Academic Press, New York, 1985, pp. 345–360.

[26] M. RENARDY, W. J. HRUSA AND J. A. NOHEL, *Mathematical Problems in Viscoelasticity*, Longman, London, 1987.

[27] P. E. ROUSE, *A theory of the linear viscoelastic properties of dilute solutions of coiling polymers*, J. Chem. Phys., 21 (1953), pp. 1271-1280.

[28] M. SLEMROD, *Instability of steady shearing flows in a nonlinear viscoelastic fluid*, Arch. Rat. Mech. Anal., 68 (1978), pp. 211-225.

[29] O. J. STAFFANS, *An inequality for positive definite Volterra kernels*, Proc. Amer. Math. Soc., 58 (1976), pp. 205-210.

[30] ———, *On a nonlinear hyperbolic Volterra equation*, this Journal, 11 (1980), pp. 793-812.

[31] B. H. ZIMM, *Dynamics of polymer molecules in dilute solutions: viscoelasticity, flow birefringence and dielectric loss*, J. Chem. Phys., 24 (1956), pp. 269-278.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

002

UNIFORM L^1 BEHAVIOR FOR THE SOLUTION OF A VOLTERRA EQUATION WITH A PARAMETER*

RICHARD NOREN†

Abstract. Consider the (scalar) initial value problem

$$\text{(P)} \qquad u_t(t,\lambda)+\lambda\int_0^t (d+a(t-\tau))u(\tau,\lambda)\,d\tau=0, \qquad u(0,\lambda)=1,$$

where $d\geqq 0$ is a constant, $\lambda\geqq 1$ is a parameter, and the subscript denotes differentiation with respect to t. The kernel $a\in L^1_{\text{loc}}[0,\infty)$ is assumed to be nonnegative, nonincreasing and convex. Let $u(t,\lambda)$ be the solution of (P) and define

$$w(t)=\sup_{\lambda\geqq 1}|u_{tt}(t,\lambda)\lambda^{-1}|.$$

Sufficient conditions (and weaker necessary conditions) concerning the kernel $a(t)$ are established in order that

$$\lim_{t\to\infty} w(t)=0 \quad \text{and} \quad \int_0^\infty w(t)\,dt<\infty.$$

Implications of the results are studied regarding solutions of the abstract initial value problem

$$\mathbf{y}'(t)+\int_0^t (d+a(t-s))\mathbf{L}\mathbf{y}(s)=\mathbf{f}(t), \qquad t>0, \quad \mathbf{y}(0)=\mathbf{x},$$

where $\mathbf{L}$ is a self-adjoint densely defined linear operator on a Hilbert space H with $\mathbf{L}\geqq\mathbf{I}$.

Key words. Volterra equation, uniform, Hilbert space, parameter, Fourier transform

AMS(MOS) subject classification. 45

1. Introduction. We study the solution $u=u(t)=u(t,\lambda)$ of the (scalar) initial value problem

$$\text{(1.1)} \qquad u'(t)+\lambda\int_0^t (d+a(t-\tau))u(\tau)\,d\tau=0, \quad u(0)=1, \quad t\geqq 0 \quad \left(' =\frac{d}{dt}\right)$$

where $\lambda\geqq 1$ is a parameter. Assuming that

(1.2) $d\geqq 0$ is a constant, $a\in L^1_{\text{loc}}[0,\infty)$ is nonnegative, nonincreasing, convex and $0=a(\infty)<a(0+)\leqq\infty$,

we give necessary conditions (and weaker sufficient conditions) concerning the kernel $a(t)$ in order that

(1.3) (i) $\displaystyle\int_0^\infty w(t)\,dt<\infty$, and

(ii) $\displaystyle\lim_{t\to\infty} w(t)=0$

hold where $w(t)\equiv\sup_{\lambda\geqq 1}|u''(t,\lambda)\lambda^{-1}|$. The necessary conditions are stated in Theorems 2.1 and 2.2. The main theorem that gives sufficient conditions is Theorem 2.5.

* Received by the editors April 7, 1986; accepted for publication (in revised form) May 12, 1987. This paper is based on the author's Ph.D. thesis, written at Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

† Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529-0077.

The parameter problem (1.1) arises in the study of the Hilbert space problem

$$y'(t)+\int_0^t (d+a(t-\tau))Ly(\tau)\,d\tau=f(t), \qquad t\geqq 0, \quad y(0)=y_0, \tag{1.4}$$

where L is a self-adjoint linear operator, defined on a dense domain D of Hilbert space H, whose spectrum is contained in $[1,\infty)$. Let

$$U(t)\equiv\int_1^\infty u(t,\lambda)\,dE_\lambda,$$

where u is the solution of (1.1) and $\{E_\lambda\}$ is the spectral family corresponding to L. The function f belongs to the class of locally Bochner integrable functions from $[0,\infty)$ to H.

Carr and Hannsgen establish the resolvent formula

$$y(t)=U(t)y_0+\int_0^t U(t-\tau)f(\tau)\,d\tau \tag{1.5}$$

for (1.4). They also give sufficient conditions in order that

$$\int_0^\infty \|U(t)\|\,dt<\infty \quad \text{and} \quad \int_1^\infty \|V(t)L^{-1/2}\|\,dt<\infty \tag{1.6}$$

hold, where $V(t)\equiv\int_1^\infty u'(t,\lambda)\,dE_\lambda$, and $\|\ \|$ denotes the operator norm for linear operators from H to H (see [2], [3]). In particular, (1.6) holds when $a(t)$ satisfies (1.2) and

$$-a' \text{ is convex.} \tag{1.7}$$

The main work in their proof is to show that the two inequalities

$$\text{(i)} \quad \int_0^\infty \sup_{\lambda\geqq 1}|u(t,\lambda)|\,dt<\infty, \text{ and} \tag{1.8}$$

$$\text{(ii)} \quad \int_0^\infty \sup_{\lambda\geqq 1}|u'(t,\lambda)\lambda^{-1/2}|\,dt<\infty$$

hold. Then (1.6) follows by the functional calculus. See [2], [3] for a discussion (with references and an example) of applications of (1.5) to problems in viscoelasticity.

The condition (1.3) implies that

$$\int_0^\infty W(t)L^{-1}\,dt<\infty \quad \text{and} \quad \lim_{t\to\infty} W(t)L^{-1}=0$$

hold where $W(t)\equiv\int_1^\infty u''(t,\lambda)\,dE_\lambda$. We show that (1.3) holds for wide classes of kernels. In particular, Theorem 2.4(ii) below shows that (1.3)(ii) holds if $a(t)$ satisfies (1.2) and (1.7), while Theorem 2.5 below implies that (1.3)(i) holds when $a(t)$ satisfies (1.2), (1.7) and one of the following: $a(0+)<\infty$, $a(t)=t^{-p}$, $0<p<1$, $a(t)=-\log t$ (small t), $a(t)=t^{-1}(-\log t)^{-q}$ (small t), $q>2$. Theorem 2.2 below shows that when (1.2) holds, then

$$\int_0^1 (-\log t)a(t)\,dt<\infty \tag{1.9}$$

is necessary for (1.3)(i) to hold. No analogous growth restriction is necessary for (1.8) to hold. We note that (1.9) rules out the locally integrable kernel $a(t)=t^{-1}(-\log t)^{-q}$ (small t), for $1<q\leqq 2$.

The simple example $a(t)=e^{-t}$ shows the need for the scaling factor $\lambda^{-1/2}$ in (1.8)(ii) and λ^{-1} in the definition of $w(t)$. In this case, (1.1) reduces to the equation $u''(t)+u'(t)+\lambda u(t)=0$, $u(0)=1$, $u'(0)=0$, and the solution is

$$u(t,\lambda)=e^{-t/2}\left(\cos \mu t+\frac{1}{2\mu}\sin \mu t\right),$$

where $\mu=(4\lambda-1)^{1/2}/2$. Differentiation shows the need to scale u', u'' by $\lambda^{-1/2}$, λ^{-1}, respectively, before taking the supremum over $\lambda \geqq 1$, in order to obtain a finite valued function of t.

Equation (1.1) is studied to establish (1.6) for more general classes of kernels in [6] and [12], with application to viscoelasticity given in [6]. The related problem of how the solution $u=u(t,\lambda)$ of (1.1) behaves as $\lambda \to \infty$ is studied in [7]. (The answer depends on whether $a'(0+)$ is finite or not. For $a(t)$ satisfying (1.2) and (1.7), $\lim_{\lambda\to\infty} u(t,\lambda)=0$ if and only if $-a'(0+)=\infty$.)

We remark that quasilinear versions of (1.4) have been under active study in recent years. See, e.g., [9] and [10] and the references therein.

In § 2 we give explicit statements of our results for (1.1). In § 3 we state applications in Hilbert space and give an example. The last two sections contain the proofs.

2. Statement of results for (1.1). Throughout this paper we assume that $d+a(t)$ satisfies (1.2). The assumptions we use to prove (1.3) involve a sufficient transform condition which then implies a (stronger) direct sufficient condition. We denote the Fourier transform of a by

$$\hat{a}(\tau)\equiv\int_0^\infty e^{-i\tau t}\,a(t)\,dt\equiv\phi(\tau)-i\tau\theta(\tau),\qquad \tau>0. \tag{2.1}$$

By Lemma 1 in [13], $\hat{a}$ is in $C^1(0,\infty)$ and by [4], ϕ and θ are nonnegative. Formally, we have

$$\hat{u}_{tt}(\tau,\lambda)=\frac{-i\tau D(\tau)}{D(\tau,\lambda)}\quad \text{for } \tau>0,\quad \lambda\geqq 1, \tag{2.2}$$

where $\hat{u}_{tt}(\tau,\lambda)\equiv\int_0^\infty e^{-i\tau t}\,u''(t,\lambda)\,dt$, and

$$D(\tau,\lambda)\equiv D(\tau)+i\tau\lambda^{-1}\equiv\hat{a}(\tau)-i\,d\tau^{-1}+i\tau\lambda^{-1}, \tag{2.3}$$

so $u''(\cdot,\lambda)$ is not in $L^1(0,\infty)$ if $D(\tau,\lambda)=0$ for some τ. By [4],

$$\phi(\tau)>0,\qquad \tau>0, \tag{2.4}$$

unless $a(t)$ is piecewise linear with changes of slope only at integral multiples of a fixed number t_0 (taken as large as possible) and τ is an integral multiple of $2\pi/t_0$. In all other cases, $D(\tau,\lambda)\neq 0$ for $\tau>0$, and Theorem 2 of [13] yields $u''(t,\lambda)$ is in $L^1(0,\infty)$ and (2.2) holds. Throughout this paper we restrict ourselves to this case by assuming (2.4). Note that (1.7) with (1.2) implies (2.4).

To give our first necessary condition for (1.3)(i), we introduce the continuous, strictly increasing function $\omega=\omega(\lambda)$, defined on some interval $[\lambda_0,\infty)$ by the formula

$$\lambda^{-1}=\theta(\omega)+d\omega^{-2}, \tag{2.5}$$

where $\omega(\lambda_0)=\rho>0$. Extend ω to $[1,\infty)$ (if $\lambda_0>1$) by defining $\omega(\lambda)=\rho$ on $[1,\lambda_0]$ (see [3]). By (2.2) and (2.5), it follows that

$$\int_0^\infty |u''(t,\lambda)|\lambda^{-1}\,dt\geqq\theta(\omega)|\hat{u}_{tt}(\omega)|\geqq\frac{\theta(\omega)\omega^2}{\lambda\phi(\omega)}\geqq\frac{\omega^2\theta^2(\omega)}{\phi(\omega)}.$$

This establishes our first result.

THEOREM 2.1. *If* (1.2) *and* (1.3)(i) *hold, then*

$$\limsup_{\tau\to\infty}\frac{(\tau\phi(\tau))^2}{\phi(\tau)}<\infty. \tag{2.6}$$

We now state our other necessary condition for (1.3)(i).

THEOREM 2.2. *If* (1.2) *and* (1.3)(i) *hold, then*

$$\int_0^1(-\log t)a(t)\,dt<\infty. \tag{2.7}$$

Define

$$C(\lambda)=\frac{[\omega^*\theta(\omega^*)]^2}{\phi(\omega^*)}, \tag{2.8}$$

where $\omega^*=\omega^*(\lambda)$ *is defined to be any number in* $[\omega/2,2\omega]$ *such that* $\phi(\omega^*)=\min_{\omega/2\leqq\tau\leqq2\omega}\phi(\tau)$.

We may now give sufficient transform conditions for (1.3)(i).

THEOREM 2.3. (i) *Suppose that* (1.2), (2.4), (2.7) *and*

$$\sup_{\lambda\geqq1}\frac{C(\lambda)}{A(\lambda/\sigma^2)}<\infty \tag{2.9}$$

are satisfied, where A *and* $\sigma=\sigma(\lambda)$ *are defined by*

$$A(x)=\int_0^x a(t)\,dt \tag{2.10}$$

and

$$\lambda^{-1}=\sigma^{-1}A(\sigma^{-1}). \tag{2.11}$$

Then $\int_0^1 w(t)\,dt<\infty$.

(ii) *Assume that* (1.2), (2.4) *and* (2.6) *are satisfied. Let* $a(t)=b(t)+c(t)$ *where* $b(t)$ *and* $c(t)$ *both satisfy the conditions stated for* $a(t)$ *in* (1.2), *except that* $b(0+)=0$ *or* $c(0+)=0$ *is permitted* (*but not both*). *Assume that*

$$\int_1^\infty\frac{b(t)}{t}\,dt<\infty \text{ and } -c' \text{ is convex.} \tag{2.12}$$

Then it follows that $\int_1^\infty w(t)\,dt<\infty$.

Combining the two parts of the above theorem and noting that (2.9) implies (2.6) yields the following result.

COROLLARY. *If* (1.2), (2.4), (2.7), (2.9) *and* (2.12) *hold, then* (1.3)(i) *holds.*

For purposes of comparison, we restate the transform conditions sufficient for (1.8)(i), (ii) to hold (see [2], [3]).

THEOREM A. *Suppose* (1.2), (2.4) *and* (2.12) *hold. Then* (1.8)(i) *holds if and only if*

$$\limsup_{\tau\to\infty}\frac{\phi(\tau)}{\phi(\tau)}<\infty. \tag{2.13}$$

THEOREM B. *Suppose* (1.2), (2.4) *and* (2.12) *hold.*

(i) *Then* (1.8)(ii) *implies that*

$$\limsup_{\tau\to\infty}\frac{\tau\theta(\tau)^{3/2}}{\phi(\tau)}<\infty. \tag{2.14}$$

(ii) *Then*

$$\limsup_{\tau\to\infty}\frac{\tau\theta(\tau)^{3/2-\varepsilon}}{\phi(\tau)}<\infty \quad \textit{for some } \varepsilon>0, \quad 0<\varepsilon<\frac{1}{2}, \tag{2.15}$$

implies that (1.8)(ii) *holds.*

Note that (2.9)→(2.6)→(2.14)→(2.13), and none of these implications may be reversed. See [3] for all but one of these facts; the fact that (2.9)→(2.6) may not be reversed is contained in [11]. Also, note that $\lim_{\lambda\to\infty}\sigma^2/\lambda = a(0+)$, by (2.11), therefore (2.6) and (2.9) are equivalent when $a(0+)<\infty$.

Although the direct conditions in Theorem 2.4(ii) and Theorem 2.5 below are slightly stronger than the transform conditions in Theorem 2.4(i) and the above corollary, they are generally much easier to check.

THEOREM 2.4. (i) *Suppose that* (1.2), (2.4) *and* (2.6) *hold. Then* (1.3)(ii) *holds.*

(ii) *Suppose that* (1.2) *and* (1.7) *hold. Then* (1.3)(ii) *holds.*

Define the functions

$$B(x)=\int_0^x -sa'(s)\,ds, \qquad x>0, \tag{2.16}$$

and

$$A_1(x)=\int_0^x sa(s)\,ds, \qquad x>0. \tag{2.17}$$

THEOREM 2.5. *Assume that* (1.2) *and* (1.7) *are satisfied.*

(a) *If* $a(0+)<\infty$, *then* (1.3)(i) *holds.*

(b) *If in addition* (2.7) *and any one of the following holds*:

(i) *There exist constants* $c_1, c_2>0$ *such that*

$$c_1\tau A_1(\tau^{-1})\leqq B(\tau^{-1})\leqq c_2\tau A_1(\tau^{-1}), \qquad \rho/2\leqq\tau, \tag{2.18}$$

(ii)

$$\lim_{\tau\to\infty}\frac{\tau A_1(\tau^{-1})}{A(\tau^{-1})}=0, \tag{2.19}$$

(iii)

$$\lim_{\tau\to\infty}\frac{B(\tau^{-1})}{A(\tau^{-1})}=0, \frac{a^2(t)}{-a'(t)} \textit{ is increasing for small } t, \textit{ and}$$

$$\int_0^{\varepsilon}\frac{a^2(t)}{-ta'(t)}\,dt<\infty \quad \textit{for some } \varepsilon>0,$$

(iv)

$$\lim_{\tau\to\infty}\frac{B(\tau^{-1})}{A(\tau^{-1})}=0 \quad \textit{and} \quad \frac{\tau A^3(\tau^{-1})}{B(\tau^{-1})}\leqq M<\infty \quad \textit{for } \rho/2\leqq\tau,$$

then (1.3)(i) *follows.*

The four cases in part (b) say roughly this about $\hat{a}(\tau)$: (i) Re $\hat{a}$ and Im $\hat{a}$ have the same order of magnitude as $\tau\to\infty$; (ii) Im $\hat{a}(\tau)$ is smaller than Re $\hat{a}(\tau)$ as $\tau\to\infty$; (iii) and (iv) Re $\hat{a}(\tau)$ is smaller than Im $\hat{a}(\tau)$ as $\tau\to\infty$. This is shown in the discussion preceding the proof of Theorem 2.5. An easy calculation shows that if $a(t)=t^{-p}$, $0<p<1$, then (b)(i) applies. If $a(t)=t^{-1}(-\log t)^{-q}$ (small t), $q>2$, then (b)(ii) applies. If $a(t)=(-\log t)$ (small t), then both (b)(iii) and (iv) apply.

The kernel

$$a(t)=\sum_{k=0}^{\infty} a_k(x_k-t)^2 X_{[0,x_k)}(t),$$

with $a_k=2^{11\times 2^{2k}}$, $x_k=2^{-4\times 2^{2k}}$, $k=0,1,2,\cdots$, is an example satisfying (1.2), (1.7) (and therefore (2.6) by Lemma 2.2(iii) of [1]) and (2.7), but neither Theorem 2.5 nor the corollary to Theorem 2.3 applies. However a modification of the proof of Theorem 2.3 shows that (1.3)(i) holds with this example (see [11]).

We close this section with the following.

CONJECTURE. Assume that (1.2), (1.7) and (2.7) are satisfied. Then (1.3)(i) holds.

3. Results in a Hilbert space. In this section, we shall see that the operator $W(t)=\int_0^\infty u''(t,\lambda)\,dE_\lambda$ may be used to solve a variant of (1.4) when (1.3)(i) holds. Then we will discuss an example of a weakly nonlinear problem where the behavior of the solution can be studied if (1.3)(i) holds.

We begin with the relation between W, V and U.

THEOREM 3.1. *Suppose* (1.2) *and* (2.6) *hold. Then for* $t>0$, $W(t)L^{-1}$ *is a bounded operator on* H *which is strongly continuous on* $(0,\infty)$. *Moreover,*

$$W(t)y=\frac{d}{dt}V(t)y=\frac{d^2}{dt^2}U(t)y, \qquad t\geqq 0, \quad y\in D.$$

Now consider the following variant of (1.4):

$$z'(t)+\int_0^t (d+a(t-s))[Lz(s)+g(s)]\,ds=f(t), \qquad t\geqq 0, \quad z(0)=z_0, \tag{3.1}$$

where $z_0\in D, f\in C([0,\infty);H)$ with $f(t)\in D(t\geqq 0)$, $Lf\in B^1_{\mathrm{loc}}([0,\infty);H)$ and $g\in B^\infty_{\mathrm{loc}}([0,\infty);D)$.

THEOREM 3.2. *Assume that* z_0, f *and* g *are as above. If* (1.3)(i) *holds, then the function* $z=z(t)$ *given by*

$$z(t)=\int_0^t W(t-s)L^{-1}G(s)\,ds \qquad \left(G(s)\equiv\int_0^s g(x)\,dx\right)$$

is the unique solution of (3.1).

The proofs of these theorems are analogous to those given in [3] and will therefore be omitted. The results of Carr and Hannsgen concerning weakly nonlinear problems also hold in the present setting. We illustrate this with a simple example.

Consider the solution $U=U(t,x)$ of

$$\begin{aligned} &U_t(x,t)+\int_0^t (d+a(t-s))(U_{xx}(s,x)+U^2(s,x)\alpha(s))\,ds=F(t,x),\\ &U(t,0)=U(t,\pi)=0, \quad t\geqq 0, \quad U(0,x)=U_0(x), \end{aligned} \tag{3.2}$$

where $\alpha\in L^1_{\mathrm{loc}}[0,\infty)$ and $\left|\int_0^t \alpha(s)\,ds\right|\leqq M<\infty$ for some M independent of $t>0$. If

$$\int_0^\pi U_0^2(x)+U_0''^2(x)\,dx<\infty$$

and one of the two quantities

$$\int_0^\infty\left\{\int_0^\pi |F(t,x)|^2\,dx\right\}^{1/2}dt, \qquad \operatorname*{ess\,sup}_{0<t<\infty}\int_0^\pi |F(t,x)|^2\,dx,$$

are sufficiently small, then (3.2) has a unique solution U such that

$$\operatorname*{ess\,sup}_{0<t<\infty} \int_0^\pi U^2(t,x)+U_{xx}^2(t,x)\,dx<\infty.$$

The details are worked out in the same way as those in the example given in [3] and therefore will be omitted here.

4. Proofs. For the remainder of the paper M will denote a positive constant whose exact value may change each time it appears.

Proof of Theorem 2.2. If we make a change of variable in the integrand of (1.1) and then differentiate the result we obtain

$$-u''(t)\lambda^{-1}=a(t)+du(t)+\int_0^t a(\tau)u'(t-\tau)\,d\tau, \qquad t>0. \tag{4.1}$$

The next lemma gives an important estimate on the integral term in (4.1).

LEMMA 4.1. *Suppose that* (1.2) *and* (2.4) *hold. Then there exist constants* $N_1, N_2>0$ *such that*

$$N_1\frac{\sigma^2}{\lambda}\leqq\sup_{t>0}\left|\int_0^t u'(t-\tau)a(\tau)\,d\tau\right|\leqq N_2\frac{\sigma^2}{\lambda}, \qquad \lambda\geqq 1. \tag{4.2}$$

The proof of the lemma relies on [3, Thm. 2.2] which says that, under the assumptions (1.4) and (2.4),

$$\frac{1}{k}\sigma\leqq\sup_{t>0}|u_t(t,\lambda)|\leqq k\sigma, \qquad \lambda\geqq 1, \tag{4.3}$$

holds for some constant k. The proof of (4.3) also contains the inequality

$$u(t,\lambda)\geqq\frac{1}{2} \quad \text{for } 0\leqq t\leqq\frac{1}{2\sigma(8+dC_2)}\equiv 2T, \tag{4.4}$$

where C_2 is a positive constant. In [5] it was shown that (1.2) implies

$$|u(t,\lambda)|\leqq 1, \qquad t\geqq 0, \quad \lambda\geqq 1. \tag{4.5}$$

If $t\leqq 1/\sigma$, use (4.2) and (2.11) to obtain

$$\left|\int_0^t u'(t-\tau)a(\tau)\,d\tau\right|\leqq M\sigma A(t)\leqq M\sigma A(\sigma^{-1})=M\sigma^2/\lambda. \tag{4.6}$$

If $1/\sigma<t$, we use (4.6), (2.11) and (4.5) to obtain

$$\begin{aligned}\left|\int_0^t u'(t-\tau)a(\tau)\,d\tau\right| &\leqq M\sigma^2/\lambda+\left|\int_{1/\sigma}^t u'(t-\tau)a(\tau)\,d\tau\right| \\ &= M\sigma^2/\lambda+\left|-a(t)+a(\sigma^{-1})u(t-\sigma^{-1})+\int_{1/\sigma}^t a'(\tau)u(t-\tau)\,d\tau\right| \\ &\leqq M\sigma^2/\lambda+2a(\sigma^{-1})\leqq M\sigma^2/\lambda.\end{aligned}$$

Taken with (4.6), we have shown that the second inequality in (4.2) holds. For the other inequality, we will need the inequality

$$2^{-3/2}A(\tau^{-1})\leqq|\hat{a}(\tau)|\leqq 4a(\tau^{-1}), \qquad \tau>0, \tag{4.7}$$

which is established in [13]. Now we let $T \leqq t \leqq 2T$, and use (1.1), (4.4) and (4.7) to obtain

$$\begin{aligned}\left|\int_0^t u'(t-\tau)a(\tau)\,d\tau\right| &= \left|\lambda \int_0^t \int_0^\tau u(\tau-s)a(s)\,ds\,a(t-\tau)\,d\tau\right| \\ &\geqq \frac{\lambda}{2}\int_{t/2}^t a(t-\tau)\int_0^t a(s)\,ds\,d\tau \\ &\geqq \frac{\lambda}{2}\left[\int_0^{T/2} a(s)\,ds\right]^2 \\ &\geqq M\lambda\left[\int_0^{1/\sigma} a(s)\,ds\right]^2 = M\sigma^2/\lambda,\end{aligned} \tag{4.8}$$

where the last inequality follows by a change of variables and (1.2). This proves the lemma.

To prove Theorem 2.2, we use (1.2), (4.5), the definition of $w(t)$ and (4.1) to observe that $\int_0^1 w(t)\,dt < \infty$ if and only if

$$\int_0^1 \sup_{\lambda \geqq 1}\left|\int_0^t a(\tau)u'(t-\tau)\,d\tau\right| dt < \infty.$$

By (4.8), it follows that

$$\begin{aligned}\left|\int_0^t a(\tau)u'(t-\tau)\,d\tau\right| &\geqq M\sigma^2/\lambda = M\sigma\int_0^{1/\sigma} a(s)\,ds \\ &\geqq \frac{M}{4(8+dC_2)t}\int_0^{2t(8+dC_2)} a(s)\,ds \\ &\geqq \frac{M}{t}\int_0^t a(s)\,ds\end{aligned}$$

for $T \leqq t \leqq 2T$. By (2.11), it follows that $\sigma \to \infty$ as $\lambda \to \infty$, and then $T \to 0$ as $\lambda \to \infty$. Therefore, for each t in $(0, \varepsilon)$ (for some $\varepsilon > 0$), there exists T with $T \leqq t \leqq 2T$. Thus the inequality $\int_0^1 w(t)\,dt < \infty$ implies that $\int_0^\varepsilon t^{-1}\int_0^t a(s)\,ds\,dt < \infty$, and then we have

$$\infty > \int_0^1 t^{-1}\int_0^t a(s)\,ds\,dt = \int_0^1 -\log sa(s)\,ds.$$

This completes the proof of Theorem 2.2.

Proof of Theorem 2.3. Except for minor details, the proof of (ii) is the same as the corresponding proofs in [2] and [3] and we will therefore not give it here. (The proof is given in [11].)

To prove (i), we will need the inequalities

$$\omega \leqq C_1\sigma, \quad \lambda \leqq C_2\sigma^2, \quad \lambda \geqq 1, \quad C_1, C_2 > 0 \quad (C_1 > 12), \tag{4.9}$$

from [3], which hold when $a(\cdot)$ satisfies (1.2). (Recall that ω is defined in (2.5).) We will also use the inequality

$$\int_0^{kx} a(s)\,ds \geqq k\int_0^x a(s)\,ds, \qquad 0 < k < 1, \quad 0 < x < \infty, \tag{4.10}$$

which follows by (1.2) and a change of variables. Finally, we will need the next lemma which will be used for Theorem 2.4 as well. We defer the proof to § 5.

LEMMA 4.2. *Under the assumptions* (1.2) *and* (2.4), *it follows that*

$$|u_{tt}(t,\lambda)\lambda^{-1}| \leqq Mt^{-1}(\sigma/\lambda + C(\lambda)) \quad \text{for } \lambda \geqq 1, \quad t > 0. \tag{4.11}$$

Now we partition $S \equiv \{(t,\lambda): t \geqq 0, \lambda \geqq 1\}$ into $S_1 \cup S_2$ where $S_1 \equiv S \cap \{(t,\lambda): \sigma^2/\lambda \leqq a(t)\}$ and $S_2 \equiv S \cap \{(t,\lambda): \sigma^2/\lambda > a(t)\}$. On S_1, we use (4.1), (4.2) and (4.5) to obtain

$$|u_{tt}(t,\lambda)\lambda^{-1}| \leqq a(t) + d + M\sigma^2/\lambda \leqq (1+M)a(t) + d \in L^1(0,1).$$

On S_2, we use (4.1), (4.2), (4.5) and (4.9) to obtain

$$|u_{tt}(t,\lambda)\lambda^{-1}| \leqq a(t) + d + M\sigma^2/\lambda \leqq (M+1+C_2 d)\sigma^2/\lambda.$$

Now partition S_2 into $S_2 = S_3 \cup S_4$ where

$$S_3 \equiv \{(t,\lambda): |u_{tt}(t,\lambda)\lambda^{-1}| \leqq t^{-1/2}\} \cap S_2,$$

and

$$S_4 \equiv \{(t,\lambda): t^{-1/2} < |u_{tt}(t,\lambda)\lambda^{-1}|\} \cap S_2.$$

On S_3, $|u_{tt}(t,\lambda)\lambda^{-1}| \leqq t^{-1/2} \in L^1(0,1)$. On S_4, we use (4.11) to observe that

$$t^{-1/2} < |u_{tt}(t,\lambda)\lambda^{-1}| \leqq \frac{M}{t}(C(\lambda) + \sigma/\lambda).$$

Thus, on S_4 we have

$$\frac{t}{\sigma/\lambda + C(\lambda)} \leqq Mt^{1/2} \quad \text{and} \quad |u_{tt}(t,\lambda)\lambda^{-1}| \leqq M\sigma^2/\lambda. \tag{4.12}$$

Define $h(x) = xA(x^{-1})$ and $g(x) = 1/A(x^{-1})$. Clearly $g(x)$ is nondecreasing. To see that $h(x)$ is nondecreasing, we write

$$h'(x) = A(x^{-1}) - x^{-1}a(x^{-1}) \geqq 0.$$

Thus, on S_4 it follows that

$$\begin{aligned} |u_{tt}(t,\lambda)\lambda^{-1}| &= h(|u_{tt}(t,\lambda)|\lambda^{-1})g(|u_{tt}(t,\lambda)|\lambda^{-1}) \\ &\leqq h\left(\frac{M}{t}(C(\lambda)+\sigma/\lambda)\right) g(M\sigma^2/\lambda) \\ &\leqq \frac{M(C(\lambda)+\sigma/\lambda)A(t^{1/2})}{A(\lambda/\sigma^2)t} \\ &\leqq \frac{M}{t}A(t^{1/2}) \in L^1(0,1), \end{aligned}$$

where the first inequality follows from (4.11), (4.12) and monotonicity, the second inequality is a consequence of (4.12) and (4.10), the last inequality follows from (2.9) and the estimate,

$$\frac{\sigma/\lambda}{A(\lambda/\sigma^2)} = \frac{A(\sigma^{-1})}{A(\sigma^{-1}\lambda/\sigma)} \leqq \frac{A(\sigma^{-1})}{A(M\sigma^{-1})} \leqq M \tag{4.13}$$

(we use (4.10) and (4.11) to obtain (4.13)) and a simple calculation using (2.7) shows that $t^{-1}A(t^{1/2}) \in L^1(0,1)$. This proves Theorem 2.3.

Proof of Theorem 2.4. Theorem 2.4(i) follows immediately from Lemma 4.2 because $\sigma/\lambda = A(\sigma^{-1})$ is bounded. In [14], Staffans proved that for $a \in L^1(0,\infty)$, (1.2) and (1.7) imply (2.6). In his thesis [1], Carr relaxed this showing that (1.2) and (1.7) imply (2.6). This result, along with Theorem 2.4(i), proves Theorem 2.4(ii).

Discussion and proof of Theorem 2.5. Before proving Theorem 2.5 we give a preliminary estimate and make some comments. To do this we will need the following inequalities: Assuming (1.2) and (1.7), it then follows that

$$CB(\tau^{-1}) \leqq \phi(\tau) \leqq KB(\tau^{-1}), \qquad \tau > 0, \tag{4.14}$$

where C, K are positive constants. Assuming (1.2), it then follows that

$$\text{(i)}\quad \tfrac{1}{5}A_1(\omega^{-1}) \leqq \lambda^{-1} \leqq C_1 A_1(\omega^{-1}), \qquad \lambda \geqq 1, \tag{4.15}$$

$$\text{(ii)}\quad \tfrac{1}{5}A_1(\tau^{-1}) \leqq \theta(\tau) \leqq 12A_1(\tau^{-1}), \qquad \tau > 0.$$

The inequalities (4.14) and (4.15) are shown in [7] and [3], respectively.

In the following we assume that (1.2) and (1.7) are satisfied. Using (4.7), (4.15) and (4.14) we obtain

$$\begin{aligned} 8^{-1}A^2(\tau^{-1}) \leqq |\hat{a}(\tau)|^2 = \phi^2(\tau) + \tau^2\theta^2(\tau) &\leqq (12B(\tau^{-1}))^2 + \tau^2(12A_1(\tau^{-1}))^2 \\ &\leqq 144(B(\tau^{-1}) + \tau A_1(\tau^{-1}))^2. \end{aligned}$$

Hence,

$$2^{-3/2}A(\tau^{-1}) \leqq 12(B(\tau^{-1}) + \tau A_1(\tau^{-1})).$$

Also, we have

$$\begin{aligned} (4A(\tau^{-1}))^2 \geqq |\hat{a}(t)|^2 = \phi^2(\tau) + \tau^2\theta^2(\tau) &\geqq (5^{-1}B(\tau^{-1}))^2 + \tau^2(5^{-1}A_1(\tau^{-1}))^2 \\ &\geqq 50^{-1}(B(\tau^{-1}) + \tau A_1(\tau^{-1}))^2. \end{aligned}$$

Hence,

$$4A(\tau^{-1}) \geqq 50^{-1/2}(B(\tau^{-1}) + \pi A_1(\tau^{-1})).$$

Combining these into one inequality yields

$$A(\tau^{-1}) \leqq (1152)^{1/2}(B(\tau^{-1}) + A_1(\tau^{-1})) \leqq 960A(\tau^{-1}). \tag{4.16}$$

In view of (4.7), (4.14) and (4.15), the behavior of $|\hat{a}(\tau)|$, $\phi(\tau) = \operatorname{Re}\hat{a}(\tau)$ and $\tau\theta(\tau) = |\operatorname{Im}\hat{a}(\tau)|$, as $\tau \to \infty$, is like that of $A(\tau^{-1})$, $B(\tau^{-1})$ and $\tau A_1(\tau^{-1})$, respectively.

In view of (4.16), condition (i) in Theorem 2.5(b) corresponds to the case where $|\hat{a}(\tau)|$, $\operatorname{Re}\hat{a}(\tau)$ and $|\operatorname{Im}\hat{a}(\tau)|$ have the same order as $\tau \to \infty$. Theorem 2.5(b)(ii) corresponds to the case where $|\operatorname{Im}\hat{a}(\tau)|$ is small compared to $|\hat{a}(\tau)|$ as $\tau \to \infty$, $|\hat{a}(\tau)|$ and $\operatorname{Re}\hat{a}(\tau)$ having the same order as $\tau \to \infty$. Theorem 2.5(b)(iii) and (iv) are both in the case where $|\operatorname{Im}\hat{a}(\tau)|$ and $|\hat{a}(\tau)|$ have the same order as $\tau \to \infty$ and $\operatorname{Re}\hat{a}(\tau)$ is small by comparison, as $\tau \to \infty$. To treat this case the additional assumptions $a^2(t)/-a'(t)$ is increasing for small t and $a^2(t)/-ta'(t) \in L^1(0,\varepsilon)$ for some ε are made in (iii), and in (iv) the extra assumption we use is $\omega A^3(\omega^{-1})/B(\omega^{-1}) < \infty$ for ω in $(\rho/2,\infty)$.

For the proof of Theorem 2.5(a), we differentiate (1.1), multiply both sides by λ^{-1} and estimate using (4.5) to obtain

$$\begin{aligned} |u_{tt}(t,\lambda)\lambda^{-1}| &\leqq (d + a(0+))|u(t,\lambda)| + \int_0^t -a'(t-\tau)|u(\tau,\lambda)|\,d\tau \\ &\leqq d + 2a(0+). \end{aligned}$$

To complete the proof, we use this bound on $(0, 1)$ and Theorem 2.3(ii) with the fact, mentioned above, that (1.2) and (1.7) imply (2.6).

Let us turn to the proof of (b). By (2.11), (4.9), (4.15) and (4.10), there exists a constant K such that

$$\frac{1}{\sigma}=\frac{1}{\lambda A(\sigma^{-1})_-}\leqq\frac{A_1(\omega^{-1})}{5A(\sigma^{-1})}\geqq\frac{KA_1(\omega^{-1})}{A(\omega^{-1})}. \tag{4.17}$$

In case (i), use (4.16), (4.17) and (2.18) to obtain

$$\frac{1}{\sigma}\geqq\frac{KA_1(\omega^{-1})}{A(\omega^{-1})}\geqq\frac{MA_1(\omega^{-1})}{B(\omega^{-1})+\omega A_1(\omega^{-1})}\geqq\frac{MA_1(\omega^{-1})}{\omega A_1(\omega^{-1})}=\frac{M}{\omega}. \tag{4.18}$$

Partition the set $S\equiv\{(t,\lambda)\colon 0\leqq t\leqq 1,\ \lambda\geqq 1\}$ into

$$S_1\equiv S\cap\{(t,\lambda)\colon t\leqq\sigma^{-1}\}\quad\text{and}\quad S_2\equiv S\cap\{(t,\lambda)\colon t>\sigma^{-1}\}.$$

For (t,λ) in S_1 use (4.1), (4.3), (4.5) and (2.7) to make the estimate

$$|u_{tt}(t,\lambda)\lambda^{-1}|\leqq a(t)+d+K\sigma\int_0^t a(\tau)\,d\tau\leqq a(t)+d+t^{-1}\int_0^t a(\tau)\,d\tau\in L^1(0,1).$$

For (t,λ) in S_2, we use (4.11), (4.14) and (4.15) to obtain

$$\begin{aligned}|u_{tt}(t,\lambda)\lambda^{-1}|&\leqq\frac{M}{t}\left(\frac{\sigma}{\lambda}+C(\lambda)\right)\leqq\frac{M}{t}\left(A(\sigma^{-1})+\frac{[\omega^*A_1(\omega^{*-1})]^2}{B(\omega^{*-1})}\right)\\&\equiv\frac{M}{t}(I_1+I_2).\end{aligned}$$

By the definition of S_2 and by (2.7), it follows that

$$t^{-1}I_1\leqq t^{-1}\int_0^t a(s)\,ds\in L^1(0,1).$$

By (4.7), (2.18), (2.7), (4.10) and (4.18), it follows that

$$\begin{aligned}t^{-1}I_2&\leqq M\omega^*A_1(\omega^{*-1})\leqq Mt^{-1}A(2\omega^{-1})\leqq Mt^{-1}A(\sigma^{-1})\\&\leqq Mt^{-1}A(t)\in L^1(0,1).\end{aligned}$$

Therefore $\int_0^1 w(t)\,dt<\infty$. With Theorem 2.3(ii), this completes the proof of Theorem 2.5(b)(ii).

To prove (b)(ii), we will show that (2.9) is satisfied and then Theorem 2.3(i) and (ii) yield (1.3)(i). Now we use (4.15) and (2.11) to write

$$\frac{\lambda}{\sigma^2}=\frac{1}{\lambda A^2(\sigma^{-1})}\geqq\frac{A_1(\omega^{-1})}{5A^2(\sigma^{-1})}=\omega^{-1}\frac{A_1(\omega^{-1})}{5\omega^{-1}A^2(\sigma^{-1})}\equiv\omega^{-1}L(\lambda).$$

There are two possibilities: $L(\lambda)\geqq 1$ or $L(\lambda)<1$.

For $L(\lambda)\geqq 1$, we have $A(\lambda/\sigma^2)\geqq A(\omega^{-1})$. Now we use (2.8), (4.14), (4.15), (4.16) and (2.19) to obtain

$$\begin{aligned}\frac{C(\lambda)}{A(\lambda/\sigma^2)}&\leqq\frac{M\omega^{*2}A_1^2(\omega^{*-1})}{A(\lambda/\sigma^2)B(\omega^{*-1})}\\&\leqq\frac{M\omega^{*2}A_1^2(\omega^{*-1})}{A(\omega^{-1})B(\omega^{*-1})}\\&=M\left(\frac{\omega^*A_1(\omega^{*-1})}{A(\omega^{*-1})}\right)\left(\frac{\omega^*A_1(\omega^{*-1})}{B(\omega^{*-1})}\right)\leqq M.\end{aligned}$$

We have shown that

$$\sup_{L(\lambda)\geqq 1}\frac{C(\lambda)}{A(\lambda/\sigma^2)}<\infty.$$

For $L(\lambda)<1$, (4.10) implies that

$$A(\lambda/\sigma^2)\geqq A(\omega^{-1}L(\lambda))\geqq L(\lambda)A(\omega^{-1});$$

then (4.16), (4.9) and (4.10) yield

$$\begin{aligned}\frac{\omega^{*2}A_1^2(\omega^{*-1})}{B(\omega^{*-1})A(\lambda/\sigma^2)}&\leqq\frac{\omega^{*2}A_1^2(\omega^{*-1})}{B(\omega^{*-1})A(\omega^{-1})L(\lambda)}\\&=\frac{5\omega^{*2}A_1^2(\omega^{*-1})\omega^{-1}A^2(\sigma^{-1})}{B(\omega^{*-1})A(\omega^{-1})A_1(\omega^{-1})}\\&\leqq\frac{M\omega A_1^2(\omega^{-1})A^2(\omega^{-1})}{B(\omega^{-1})A(\omega^{-1})A_1(\omega^{-1})}\\&\leqq M\omega A_1(\omega^{-1})\leqq M,\end{aligned}$$

where we have used the following easily obtained inequalities:

$$A_1(2x)\leqq 4A_1(x)\quad\text{and}\quad B(2x)\leqq 4B(x),\quad x>0. \tag{4.19}$$

This shows that

$$\sup_{L(\lambda)<1}\frac{C(\lambda)}{A(\lambda/\sigma^2)}<\infty,$$

which establishes (2.9) and finishes the proof.

To prove (iii) we use (4.16), (4.17) and the assumption of (iii) to obtain

$$\sigma^{-1}\geqq\frac{KA_1(\omega^{-1})}{A(\omega^{-1})}\geqq\frac{M\omega^{-1}A(\omega^{-1})}{A(\omega^{-1})}=M\omega^{-1}. \tag{4.20}$$

Thus $\sigma^{-1}\geqq M\omega^{-1}$ as in (4.18) and the rest of the proof follows exactly as in the lines following (4.18) except for the term $t^{-1}I_2$ which we shall treat next. By (4.19) and integration by parts, we have

$$\begin{aligned}t^{-1}I_2&\leqq Mt^{-1}\frac{\omega^{*2}A_1^2(\omega^{*-1})}{B(\omega^{-1})}\leqq Mt^{-1}\frac{\omega^2A_1^2(\omega^{-1})}{B(\omega^{-1})}\\&=\frac{Mt^{-1}\omega^2}{2B(\omega^{-1})}\left(\omega^{-2}a(\omega^{-1})+\int_0^{\omega^{-1}}-s^2a'(s)\,ds\right)^2\\&\leqq Mt^{-1}\left(\frac{a^2(\omega^{-1})}{\omega^2B(\omega^{-1})}+\frac{\omega^2\left(\int_0^{\omega^{-1}}-s^2a'(s)\,ds\right)^2}{B(\omega^{-1})}\right)\\&\equiv Mt^{-1}(J_1+J_2).\end{aligned}$$

By definition of S_2, (4.20) and assumption (iii), we have

$$\begin{aligned}t^{-1}J_1&\leqq M\frac{a^2(\omega^{-1})}{-ta'(\omega^{-1})}\\&\leqq\frac{Ma^2(M_1t)}{-ta'(M_1t)}\in L^1\left(0,\frac{\varepsilon}{M_1}\right)\quad(\text{some } M_1>0).\end{aligned}$$

Also,

$$\begin{aligned} t^{-1}J_2 &\leqq \frac{M\omega^2 \int_0^{\omega^{-1}} -sa'(s)\,ds \int_0^{\omega^{-1}} -s^3a'(s)\,ds}{tB(\omega^{-1})} \\ &= \frac{M\omega^2 \int_0^{\omega^{-1}} -s^3a'(s)\,ds}{t} \\ &\leqq Mt^{-1}\int_0^{\omega^{-1}} -sa'(s)\,ds \\ &\leqq Mt^{-1}\int_0^{t} -sa'(s)\,ds \\ &\leqq Mt^{-1}A(t) \in L^1(0,1), \end{aligned}$$

where we have used the Cauchy-Schwarz inequality, the definition of S_2 (4.20), (4.19) and (2.7). Note that on $[\varepsilon/M_1, 1]$, (4.1) can be used to show that

$$|u_{tt}(t,\lambda)\lambda^{-1}| \leqq a(\varepsilon/M_1) + d + 2a(\varepsilon/M_1) \leqq M.$$

Together with Theorem 2.3, this finishes the proof of (iii).

To prove (iv), we will show that (2.9) holds and then Theorem 2.3 applies, finishing the proof. To do this use (4.15), (2.11), (4.9), (4.14), (4.16), (4.10) and (4.19) to obtain

$$\begin{aligned} \frac{\omega^{*2}\theta^2(\omega^*)}{\theta(\omega^*)A(\lambda/\sigma^2)} &\leqq \frac{\omega^{*2}A_1^2(\omega^{*-1})}{B(\omega^{*-1})A(\lambda/\sigma^2)} \leqq \frac{MA^2(\omega^{*-1})}{B(\omega^{*-1})A(\lambda/\sigma^2)} \\ &\leqq \frac{MA^2(\omega^{-1})}{B(\omega^{-1})\lambda/\sigma^2 A(1)} \leqq \frac{MA^2(\omega^{-1})A^2(\sigma^{-1})}{B(\omega^{-1})A_1(\omega^{-1})} \\ &\leqq \frac{MA^4(\omega^{-1})}{B(\omega^{-1})A_1(\omega^{-1})} \leqq \frac{M\omega A^3(\omega^{-1})}{B(\omega^{-1})} \\ &\leqq \frac{M\omega A^3(\omega^{-1})}{B(\omega^{-1})} \leqq M. \end{aligned}$$

This finishes the proof of Theorem 2.5.

5. Proof of Lemma 4.2. When (1.2) and (2.4) hold, we have the inversion formula

$$\pi u_{tt}(t,\lambda) = \operatorname{Re}\frac{1}{t\lambda}\int_0^\infty e^{i\tau t}\left(\frac{i\tau^2 D'(\tau) + \lambda D(\tau)^2}{D(\tau,\lambda)^2}\right) d\tau, \qquad t \geqq 0, \quad \lambda \geqq 1, \tag{5.1}$$

where the integral is absolutely convergent at both $\tau = 0$ and $\tau = \infty$. This was established in [11]. Thus, we have that

$$|u_{tt}(t,\lambda)\lambda^{-1}| \leqq \frac{1}{t\lambda^2}\int_0^\infty \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| + \left|\frac{\lambda D(\tau)^2}{D(\tau,\lambda)^2}\right| d\tau. \tag{5.2}$$

But, (5.19) and (5.21) of [3] and Lemma 1 of [13] imply that

$$|D(\tau,\lambda)| \geqq \max\left\{\phi(\tau), \frac{d-\tau^2}{\tau}\right\} \geqq \frac{1}{M\tau}, \qquad 0 < \tau \leqq \rho, \quad d > 0, \tag{5.3}$$

$$|D(\tau,\lambda)| \geqq \max\{2^{-3/2}A(\tau^{-1}) - \tau, \phi(\tau)\} \geqq M, \qquad 0 < \tau \leqq \rho, \quad d = 0, \tag{5.4}$$

and

$$|\hat{a}'(\tau)| \leqq 40A_1(\tau^{-1}), \qquad \tau > 0. \tag{5.5}$$

We use (5.3) and (5.5) to obtain (for $d > 0$)

$$\begin{aligned}\int_0^\rho \left|\frac{D(\tau)}{D(\tau,\lambda)}\right|^2 d\tau &= \int_0^\rho \left|1-\frac{i\tau}{\lambda D(\tau,\lambda)}\right|^2 d\tau \\ &\leqq \int_0^\rho 2+\frac{2\tau^2}{\lambda^2|D(\tau,\lambda)|^2}\, d\tau \\ &\leqq 2\rho+\frac{M}{\lambda^2}\int_0^\rho \frac{\tau^2}{\tau^{-2}}\, d\tau \leqq M\end{aligned}$$

and

$$\int_0^\rho \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| d\tau \leqq M\int_0^\rho \frac{\tau^2(\tau^{-2})}{\tau^{-2}}\, d\tau \leqq M.$$

When $d = 0$, we use (5.4) to obtain

$$\int_0^\rho \left|\frac{D(\tau)}{D(\tau,\lambda)}\right|^2 d\tau \leqq \int_0^\rho 2+\frac{2\tau^2}{\lambda^2|D(\tau,\lambda)|^2}\, d\tau \leqq 2\rho+\frac{M}{\lambda^2}\int_0^\rho \tau^2\, d\tau \leqq M.$$

Also,

$$\int_0^\rho \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| d\tau \leqq M\int_0^\rho \frac{\tau A(\tau^{-1})\, d\tau}{(\max\{2^{-3/2}A(\tau^{-1})-\tau,\, \phi(\tau)\})^2} \leqq M,$$

where we have used

$$\lim_{\tau\to 0} \tau A(\tau^{-1}) = 0$$

and

$$0 < 8^{-1}\left(\int_0^\infty a(s)\, ds\right)^2 = \lim_{\tau\to 0+} (2^{-3/2}A(\tau^{-1})-\tau)^2 \leqq \infty.$$

Therefore, we have

$$\frac{1}{\lambda^2 t}\int_0^\rho \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| + \left|\frac{\lambda D^2(\tau)}{D(\tau,\lambda)^2}\right| d\tau \leqq Mt^{-1}\sigma/\lambda. \tag{5.6}$$

For our estimates on (ρ, ∞) we will need the inequalities

$$\tau\lambda^{-1} \leqq M|D(\tau,\lambda)|, \qquad 2\omega \leqq \tau < \infty, \tag{5.7}$$

$$A(\tau^{-1}) \leqq M|D(\tau,\lambda)|, \qquad \rho/2 \leqq \tau \leqq \omega/2, \quad 2\omega \leqq \tau < \infty, \tag{5.8}$$

$$\lambda \leqq M\omega^2, \qquad \lambda \geqq 1, \tag{5.9}$$

$$MA(\tau^{-1}) \leqq \phi(\omega^*)+\omega^*\theta(\omega^*), \qquad \omega/2 \leqq \tau \leqq 2\omega, \tag{5.10}$$

$$\phi^2(\tau)+\frac{|\tau-\omega|^2}{\lambda^2} \leqq M|D(\tau,\lambda)|^2, \qquad \tau \leqq \rho/2, \tag{5.11}$$

which are contained in [3].

First we use (5.7), (2.11), (4.10), (5.5) and the Fubini theorem to obtain

$$\int_{2C_1\sigma}^{\infty} \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| d\tau \leqq \lambda^2 M \int_{2C_1\sigma}^{\infty} (d\tau^{-2} + A_1(\tau^{-1}))\, d\tau$$

$$\leqq \lambda^2 M \left(\sigma^{-1} + \int_{2C_1\sigma}^{\infty}\int_0^{1/\tau} sa(s)\, ds\, d\tau\right)$$

$$\leqq M\lambda^2 \left(\sigma/\lambda + \int_0^{1/2C_1\sigma} a(s)\, ds\right) \leqq M\lambda\sigma.$$

Also, by (5.7), (4.10) and (4.7), it follows that

$$\int_{2C_1\sigma}^{\infty} \left|\frac{D(\tau)}{D(\tau,\lambda)}\right|^2 d\tau \leqq M\lambda^2 \int_{2C_1\sigma}^{\infty} \frac{A^2(\tau^{-1})}{\tau^2}\, d\tau$$

$$\leqq M\lambda^2 \int_{2C_1\sigma}^{\infty} \tau^{-2}\, d\tau A^2(\sigma^{-1}/2C_1)$$

$$\leqq M\lambda^2 \sigma^{-1} A^2(\sigma^{-1}) = M\sigma.$$

Therefore, we have

$$\frac{1}{t\lambda^2}\int_{2C_1\sigma}^{\infty} \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| + \left|\frac{\lambda D(\tau)^2}{D(\tau,\lambda)^2}\right| d\tau \leqq \frac{M\sigma}{t\lambda}. \tag{5.12}$$

Next we use (5.8), (5.5), (4.10) and (2.11) to obtain

$$\lambda^{-2}\left(\int_\rho^{\omega/2} + \int_{2\omega}^{2C_1\sigma}\right) \left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right| d\tau \leqq M\lambda^{-2} \int_\rho^{2C_1\sigma} \frac{\tau^2(d\tau^{-2} + A_1(\tau^{-1}))}{A(\tau^{-1})^2}\, d\tau$$

$$\leqq M\lambda^{-2} \int_\rho^{2C_1\sigma} A^{-2}(\tau^{-1}) + \tau A^{-1}(\tau^{-1})\, d\tau$$

$$\leqq M\lambda^{-2}(\sigma A^{-2}(1/2C_1\sigma) + \sigma^{-2}A^{-1}(1/2C_1\sigma))$$

$$\leqq M\lambda^{-2}(\sigma A^{-2}(\sigma^{-1}) + \sigma^2 A^{-1}(\sigma^{-1}))$$

$$= M(\sigma^{-1} + \sigma/\lambda) \leqq M\sigma/\lambda,$$

where the last inequality is due to the fact that $0 < a(0+) = \lim_{\sigma\to\infty} \sigma^2/\lambda \leqq \infty$. Also, by (5.8), (4.9) and monotonicity of the function $h(x) = x/A(x^{-1})$, we have

$$\lambda^{-1}\int_\rho^{\omega/2} \left|\frac{D(\tau)}{d(\tau,\lambda)}\right|^2 d\tau \leqq \lambda^{-1}\int_\rho^{\omega/2} 2 + \frac{2\tau^2}{\lambda^2|D(\tau,\lambda)|^2}\, d\tau$$

$$\leqq \omega/\lambda + M\lambda^{-3}\int_\rho^{\omega/2} \frac{\tau^2}{A^2(\tau^{-1})}\, d\tau$$

$$\leqq M\left(\frac{\sigma}{\lambda} + \frac{\omega\sigma^2}{\lambda^3 A^2(1/2C_1\sigma)}\right)$$

$$\leqq M\left(\frac{\sigma}{\lambda} + \frac{\sigma^3}{\lambda^3 A^2(\sigma^{-1})}\right) \leqq M\sigma/\lambda.$$

Also, by (4.7), (5.8), (5.9), and (4.10), it follows that

$$\lambda^{-1}\int_{2\omega}^{2C_1\sigma}\left|\frac{D(\tau)}{D(\tau,\lambda)}\right|^2 d\tau \leqq M\lambda^{-1}\int_{2\omega}^{2C_1\sigma}\frac{(A(\tau^{-1})+d\tau^{-2})^2}{A^2(\tau^{-1})}\,d\tau$$

$$= M\lambda^{-1}\int_{2\omega}^{2C_1\sigma}\left(1+\frac{d}{\tau^2 a(\tau^{-1})}\right)^2 d\tau$$

$$\leqq M\frac{\sigma}{\lambda}\left(1+\frac{1}{\omega^2 A(\sigma^{-1})}\right)^2$$

$$\leqq M\frac{\sigma}{\lambda}(1+\sigma^{-1})^2 \leqq M\frac{\sigma}{\lambda}.$$

Thus, we have the estimate

$$\frac{1}{t\lambda^2}\left(\int_{\rho/2}^{\omega/2}+\int_{2\omega}^{2C_1\sigma}\right)\left(\left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right|+\left|\frac{\lambda D(\tau)^2}{D(\tau,\lambda)^2}\right|\right)d\tau \leqq \frac{M}{t}\frac{\sigma}{\lambda}. \tag{5.13}$$

On the interval $(\omega/2, 2\omega)$, we use (5.11), (5.5), (4.9), (5.10) and (4.15) to obtain

$$\lambda^{-2}\int_{\omega/2}^{2\omega}\left|\frac{\tau^2 D'(\tau)}{D(\tau,\lambda)^2}\right|d\tau \leqq M\lambda^{-2}\int_{\omega/2}^{2\omega}\frac{\tau A(\tau^{-1})\,d\tau}{\phi^2(\tau)+((\tau-\omega)/\lambda)^2}$$

$$\leqq M\int_{\omega/2}^{2\omega}\frac{\tau\,d\tau}{(\phi(\omega^*)\lambda)^2+|\tau-\omega|^2}A(2\omega^{-1})$$

$$\leqq M\int_{\omega/2}^{2\omega}\frac{\tau\,d\tau}{(\phi(\omega^*)\lambda)^2+|\tau-\omega|^2}A(2\omega^{-1})$$

$$\leqq M\omega^*\int_{\omega/2}^{2\omega}\frac{d\tau}{(\phi(\omega^*)\lambda)^2+|\tau-\omega|^2}(\phi(\omega^*)+\omega^*\theta(\omega^*))$$

$$\leqq M\frac{\omega^*(\phi(\omega^*)+\omega^*\theta(\omega^*))}{\lambda\phi(\omega^*)}$$

$$\leqq M\left(\frac{\sigma}{\lambda}+\frac{(\omega^*\theta(\omega^*))^2}{\phi(\omega^*)}\right).$$

Also, by (4.9), (5.11) and (4.15), we may write

$$\lambda^{-1}\int_{\omega/2}^{2\omega}\left|\frac{D(\tau)}{D(\tau,\lambda)}\right|^2 d\tau = \lambda^{-1}\int_{\omega/2}^{2\omega}\left|1-\frac{i\tau\lambda^{-1}}{D(\tau,\lambda)}\right|^2$$

$$\leqq 2\lambda^{-1}\int_{\omega/2}^{2\omega}1+\frac{\tau^2\lambda^{-2}}{|D(\tau,\lambda)|^2}\,d\tau$$

$$\leqq M\lambda^{-1}\int_{\omega/2}^{2\omega}1+\frac{\tau^2\lambda^{-2}}{\phi^2(\tau)+|(\tau-\omega)/\lambda|^2}\,d\tau$$

$$\leqq M\lambda^{-1}\int_{\omega/2}^{2\omega}1+\frac{\tau^2}{(\lambda\phi(\omega^*))^2+|\tau-\omega|^2}\,d\tau$$

$$\leqq M\left(\frac{\omega}{\lambda}+\frac{\omega^2}{\phi(\omega^*)\lambda^2}\right)$$

$$\leqq M\left(\frac{\sigma}{\lambda}+\frac{(\omega^*\theta(\omega^*))^2}{\phi(\omega^*)}\right).$$

These two estimates on $(\omega/2, 2\omega)$ together with (5.13), (5.12), (5.6) and (5.2) yield (4.11) (see (2.8)), thus proving Lemma 4.2.

Acknowledgment. The author thanks Professor Kenneth B. Hannsgen for many enlightening discussions on this material.

REFERENCES

[1] R. W. CARR, *Uniform L^p estimates for a linear integrodifferential equation with a parameter*, Ph.D. thesis, University of Wisconsin-Madison, Madison, WI, 1977.

[2] R. W. CARR AND K. B. HANNSGEN, *A nonhomogeneous integrodifferential equation in Hilbert space*, this Journal, 10 (1979), pp. 961-984.

[3] ———, *Resolvent formulas for a Volterra equation in Hilbert space*, this Journal, 13 (1982), pp. 459-483.

[4] K. B. HANNSGEN, *Indirect Abelian theorems and a linear Volterra equation*, Trans. Amer. Math. Soc., 142 (1969), pp. 539-555.

[5] ———, *A Volterra equation with parameter*, this Journal, 4 (1973), pp. 22-30.

[6] ———, *A linear integrodifferential equation for viscoelastic rods and plates*, Quart. Appl. Math., 41 (1983), pp. 75-84.

[7] K. B. HANNSGEN AND R. L. WHEELER, *Behavior of the solution of a Volterra equation as a parameter tends to infinity*, J. Integral Equations, 7 (1984), pp. 229-237.

[8] ———, *Uniform L^1 behavior in classes of integrodifferential equations with completely montonic kernels*, this Journal, 15 (1984), pp. 579-594.

[9] W. J. HRUSA AND J. A. NOHEL, *Global existence and asymptotics in one-dimensional nonlinear viscoelasticity*, Proc. 5th Symposium on Trends in Applications of Pure Mathematics to Mechanics, Springer Lecture Notes in Physics 195, 1984, pp. 165-187.

[10] W. J. HRUSA AND M. RENARDY, *On a class of quasilinear partial integrodifferential equations with singular kernels*, J. Differential Equations, to appear.

[11] R. D. NOREN, *Uniform L^1 behavior for the solution of a Volterra equation with a parameter*, Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1985.

[12] ———, *A linear Volterra integrodifferential equation for viscoelastic rods and plates*, Quart. Appl. Math., submitted.

[13] D. F. SHEA AND S. WAINGER, *Variants of the Wiener-Levy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312-343.

[14] O. J. STAFFANS, *An inequality for positive definite Volterra kernels*, Proc. Amer. Math. Soc., 58 (1976), pp. 205-210.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

003

HOMOGENIZATION OF STATIONARY FLOW OF MISCIBLE FLUIDS IN A DOMAIN WITH A GRAINED BOUNDARY*

A. MIKELIĆ† AND I. AGANOVIĆ‡

Abstract. The purpose of this paper is to obtain phenomenological equations for stationary miscible displacement in a domain with a grained boundary (porous medium), with the help of homogenization of fluid mechanics equations. It is assumed that the viscosity of the mixture depends on the concentration of the solvent.

Key words. homogenization, miscible flow, porous medium

AMS(MOS) subject classifications. 35B40, 76S05

1. Introduction. In this paper we consider the homogenization of fluid mechanics equations of stationary miscible flow through a porous medium, assuming that the viscosity of the mixture depends on the concentration of the solvent.

We use the standard notation required for the homogenization method. Let $Y=]0,1[^n$, $n=2$ or 3; let $\mathcal{O}\subset Y$ be an open set strictly included in Y and locally placed on one side of its boundary $S\in C^\infty$ and $Y^*=Y\setminus\bar{\mathcal{O}}$. For $k\in Z^n$, we define $Y_k=Y+k$, $\mathcal{O}_k=\mathcal{O}+k$. Let $\Omega\subset\mathbf{R}^n$ be a bounded domain locally placed on one side of its boundary $\Gamma\in C^2$. For sufficiently small $\varepsilon>0$, we consider the sets

$$T_\varepsilon=\{k\in\mathbf{Z}^n:\ \varepsilon Y_k\subset\Omega\},\qquad K_\varepsilon=\{k\in\mathbf{Z}^n:\ \varepsilon Y_k\cap\Gamma\neq\varnothing\},$$

and define

$$\mathcal{O}_\varepsilon=\bigcup_{k\in T_\varepsilon}\varepsilon\mathcal{O}_k,\ S_\varepsilon=\partial\mathcal{O}_\varepsilon,\ \Omega_\varepsilon=\Omega\setminus\bar{\mathcal{O}}_\varepsilon.$$

Obviously, $\partial\Omega_\varepsilon=\Gamma\cup S_\varepsilon$. The domains $\mathcal{O}_\varepsilon$ and Ω_ε represent, respectively, the solid and fluid part of the porous medium Ω. We consider the Stokes flow of a mixture of two incompressible fluids in the domain Ω_ε, upon stationary diffusion of one of them (solvent) into the other.

Notation. v^ε, p^ε = the velocity and the pressure of the mixture, respectively; s^ε = the concentration of the solvent; $\varepsilon^2\mu(s^\varepsilon)$ = the viscosity of the mixture; f = the density of the external body force; $d=\text{const}>0$ = the diffusion coefficient of the mixture.

Taking into account the simplest model of diffusion, for the velocity, the pressure and the concentration we have the following equations and boundary conditions (e.g., [6]):

$$-\nabla p^\varepsilon+\varepsilon^2\operatorname{div}(\mu(s^\varepsilon)\nabla v^\varepsilon)+f=0\quad\text{in }\Omega_\varepsilon, \tag{1.1}$$

$$\operatorname{div} v^\varepsilon=0\quad\text{in }\Omega_\varepsilon, \tag{1.2}$$

$$v^\varepsilon=h\quad\text{on }\Gamma,\qquad v^\varepsilon=0\quad\text{on }S_\varepsilon, \tag{1.3}$$

$$-d\Delta s^\varepsilon+v^\varepsilon\cdot\nabla s^\varepsilon=0\quad\text{in }\Omega_\varepsilon, \tag{1.4}$$

$$s^\varepsilon=g\quad\text{on }\Gamma,\qquad\frac{\partial s^\varepsilon}{\partial\nu}=0\quad\text{on }S_\varepsilon. \tag{1.5}$$

* Received by the editors December 4, 1985; accepted for publication (in revised form) April 9, 1987. This work was supported in part by INA-NAFTAPLIN, Geological Exploration and Development Division, Zagreb, Yugoslavia. The results were presented at the SIAM Spring Meeting held in Pittsburgh, Pennsylvania, June 1985.

† Ruder Bošković Institute, 41001 Zagreb, Yugoslavia.

‡ Department of Mathematics, University of Zagreb, 41001 Zagreb, Yugoslavia.

Here h and g are given functions and ν denotes the unit vector of the outward normal on the boundary of a domain. We assume that

$$\mu \in C^1([0,1]), \quad \mu > 0 \quad \text{on } [0,1], \tag{1.6}$$

$$f \in (C(\bar{\Omega}))^n, \tag{1.7}$$

$$h \in (C^2(\bar{\Omega}))^n, \qquad \int_\Gamma h \cdot \nu \, d\Gamma = 0, \tag{1.8}$$

$$g \in C^2(\bar{\Omega}), \qquad 0 \leqq g \leqq 1 \text{ on } \bar{\Omega}, \tag{1.9}$$

and consider μ extended to $\mathbf{R}$ as follows:

$$\mu(s) = \begin{cases} \mu(0), & s < 0, \\ \mu(1), & s > 1. \end{cases}$$

By use of the Leray-Schauder Fixed Point Theorem and the Maximum Principle, one can prove the following result.

THEOREM 1.1. *Under the assumptions* (1.6)-(1.9), *the problem* (1.1)-(1.5) *has at least one solution*

$$(v^\varepsilon, p^\varepsilon, s^\varepsilon) \in (H^1(\Omega_\varepsilon))^n \times (L^2(\Omega_\varepsilon)/\mathbf{R}) \times (H^1(\Omega_\varepsilon) \cap C(\bar{\Omega}_\varepsilon)); \tag{1.10}$$

the inequality

$$0 \leqq s^\varepsilon \leqq 1 \quad \text{a.e. in } \Omega_\varepsilon \tag{1.11}$$

holds true for each solution.

We shall prove that the limit of solutions (1.10) (as ε tends to zero) satisfies the equations and boundary conditions of the simplest phenomenological model of miscible flow through a porous medium (e.g., [3]).

2. Macroscopic and constitutive equations. In this section, $C > 0$ denotes a generic constant which does not depend on ε and has possibly different values at different places.

LEMMA 2.1. *Let* $(v^\varepsilon, p^\varepsilon, s^\varepsilon)$ *be a solution to problem* (1.1)-(1.5), *and*

$$\tilde{v}^\varepsilon = \begin{cases} v^\varepsilon & \text{in } \Omega_\varepsilon, \\ 0 & \text{in } \mathcal{O}_\varepsilon. \end{cases}$$

Then

$$\|\nabla \tilde{v}^\varepsilon\|_{(L^2(\Omega))^{n^2}} \leqq \frac{C}{\varepsilon},$$

$$\|\tilde{v}^\varepsilon\|_{(L^2(\Omega))^n} \leqq C.$$

There exists the extension $\tilde{p}^\varepsilon \in L^2(\Omega)/\mathbf{R}$ *of the function* p^ε, *such that*

$$\|\tilde{p}^\varepsilon\|_{L^2(\Omega)} \leqq C.$$

LEMMA 2.2. *Under the notation of the preceding lemma, there exist subsequences of* $\{\tilde{v}^\varepsilon\}$ *and* $\{\tilde{p}^\varepsilon\}$ (*denoted again by* $\{\tilde{v}^\varepsilon\}, \{\tilde{p}^\varepsilon\}$) *and functions* $v \in (L^2(\Omega))^n$, $p \in L^2(\Omega)/\mathbf{R}$, *such that*

$$\tilde{v}^\varepsilon \to v \quad \text{weakly in } L^2(\Omega),$$

$$\tilde{p}^\varepsilon \to p \quad \text{strongly in } L^2(\Omega)/\mathbf{R},$$

as $\varepsilon \to 0$. *The function* v *satisfies the following* (*macroscopic*) *equation and boundary condition*:

$$\operatorname{div} v = 0 \quad \text{in } \Omega,$$

$$v \cdot h = \nu \cdot h \quad \text{on } \Gamma.$$

LEMMA 2.3. *Under the notation of Lemma* 2.1, *there exist the extension* $\tilde{s}^\varepsilon \in H^1(\Omega) \cap L^\infty(\Omega)$ *of the function* s^ε, *such that*

$$\|\tilde{s}^\varepsilon\|_{H^1(\Omega)} \leqq C,$$

$$\|\tilde{s}^\varepsilon\|_{L^\infty(\Omega)} \leqq C.$$

Let

$$\tilde{\sigma}^\varepsilon = \begin{cases} \sigma^\varepsilon = d\nabla s^\varepsilon & \text{in } \Omega_\varepsilon, \\ 0 & \text{in } \mathcal{O}_\varepsilon. \end{cases}$$

Then

$$\|\tilde{\sigma}^\varepsilon\|_{(L^2(\Omega))^n} \leqq C,$$

$$\operatorname{div} \tilde{\sigma}^\varepsilon - \tilde{v}^\varepsilon \cdot \nabla \tilde{s}^\varepsilon = 0 \quad \text{in } \Omega.$$

LEMMA 2.4. *Under the notation of the preceding lemma, there exist subsequences of* $\{\tilde{s}^\varepsilon\}$, $\{\tilde{\sigma}^\varepsilon\}$ (*denoted again by* $\{\tilde{s}^\varepsilon\}$, $\{\tilde{\sigma}^\varepsilon\}$) *and functions* $s \in H^1(\Omega) \cap L^\infty(\Omega)$, $\sigma \in (L^2(\Omega))^n$, *such that*

$$\tilde{s}^\varepsilon \to s \quad \text{weakly in } H^1(\Omega) \text{ and weakly* in } L^\infty(\Omega),$$

$$\tilde{\sigma}^\varepsilon \to \sigma \quad \text{weakly in } L^2(\Omega),$$

as $\varepsilon \to 0$. *The functions* s *and* σ *satisfy the following* (*macroscopic*) *equation and boundary condition*:

$$\operatorname{div} \sigma - v \cdot \nabla s = 0 \quad \text{in } \Omega,$$

$$s = g \quad \text{on } \Gamma.$$

Let $q \in (H^1(Y^*))^n/\mathbf{R}$ *be the solution to the problem*

$$\Delta q = 0 \quad \text{in } Y^*,$$

$$(\nabla q + I)\nu = 0 \quad \text{on } S,$$

q *is* Y*-periodic*

(*where* I *denotes the unity matix*), *and*

$$A = \int_{Y^*} \nabla q \, dy, \quad \theta = \operatorname{meas} Y^*,$$

$$D = d(\theta I + A).$$

LEMMA 2.5. *The functions* s *and* σ, *defined by Lemma* 2.4, *satisfy the following* (*constitutive*) *equation*:

$$\sigma = D\nabla s.$$

The proofs of Lemmas 2.1-2.5 can be performed by adapting the method of L. Tartar, applied in the linear case (see [8, Appendix], [4] and [7]).

Let $(w^i, \pi^i) \in (H^1(Y^*))^n \times (L^2(Y^*)/\mathbf{R})(i = 1, \cdots, n)$ be the solution to the problem

$$-\nabla \pi^i + \Delta w^i + e^i = 0 \quad \text{in } Y^*, \tag{2.1}$$

$$\operatorname{div} w^i = 0 \quad \text{in } Y^*, \tag{2.2}$$

$$w^i = 0 \quad \text{on } S, \tag{2.3}$$

$$w^i, \pi^i \quad \text{are } Y\text{-periodic} \tag{2.4}$$

and

$$K = (K_{ij}), \qquad K_{ij} = \int_{Y^*} (w^i)_j \, dy.$$

In the sequel, we consider w^i extended by zero to $\mathcal{O}$.

LEMMA 2.6. *The functions v, p and s, defined by Lemmas* 2.2 *and* 2.4, *satisfy the following* (*constitutive*) *equation* (*Darcy's law*):

$$v = \frac{1}{\mu(s)} K(f - \nabla p). \tag{2.5}$$

Proof. Let

$$w^{i,\varepsilon}(x) = w^i\left(\frac{x}{\varepsilon}\right), \quad \pi^{i,\varepsilon}(x) = \pi^i\left(\frac{x}{\varepsilon}\right), \quad i = 1, \cdots, n.$$

As a consequence of the basic lemma on periodic extension [8, p. 57], we obtain that

$$(w^{i,\varepsilon})_j \to K_{ij} \quad \text{weakly in } L^2(\Omega), \tag{2.6}$$

as $\varepsilon \to 0$. Using the regularity of the functions w^i and π^i we obtain the inequalities

$$\|w^{i,\varepsilon}\|_{(L^\infty(\Omega_\varepsilon))^n} \leqq C, \tag{2.7}$$

$$\|\nabla w^{i,\varepsilon}\|_{(L^\infty(\Omega_\varepsilon))^{n^2}} \leqq \frac{C}{\varepsilon}, \tag{2.8}$$

$$\|\pi^{i,\varepsilon}\|_{L^\infty(\Omega_\varepsilon)} \leqq C. \tag{2.9}$$

(Here we consider $\pi^{i,\varepsilon}$ extended in a simple way to $\varepsilon\mathcal{O}_{-k}$, $k \in K_\varepsilon$.) The functions $w^{i,\varepsilon}$, $\pi^{i,\varepsilon}$ satisfy the equations

$$-\varepsilon \nabla \pi^{i,\varepsilon} + \varepsilon^2 \Delta w^{i,\varepsilon} + e^i = 0 \quad \text{in } \Omega_\varepsilon \setminus \bigcup_{k \in K_\varepsilon} \varepsilon \bar{\mathcal{O}}_k, \tag{2.10}$$

$$\operatorname{div} w^{i,\varepsilon} = 0 \quad \text{in } \Omega_\varepsilon \setminus \bigcup_{k \in K_\varepsilon} \varepsilon \bar{\mathcal{O}}_k. \tag{2.11}$$

Under the notation of Lemma 2.1 and Lemma 2.3, for $\varphi \in \mathscr{D}(\Omega)$ we have (because of (1.1) and (2.10))

$$\int_\Omega (-\nabla \tilde{p}^\varepsilon + \varepsilon^2 \operatorname{div}(\mu(\tilde{s}^\varepsilon)\nabla \tilde{v}^\varepsilon) + f) \cdot w^{i,\varepsilon}\varphi \, dx = 0,$$

$$\int_\Omega \mu(\tilde{s}^\varepsilon)(-\varepsilon \nabla \pi^{i,\varepsilon} + \varepsilon^2 \Delta w^{i,\varepsilon} + e^i) \cdot \tilde{v}^\varepsilon \varphi \, dx = 0.$$

(Here we have assumed that ε is sufficiently small, so that $\operatorname{supp} \varphi \cap \varepsilon \mathscr{O}_k = \varnothing$ for $k \in K_\varepsilon$.) Subtracting these equations and performing the integation by parts (and using (1.2), (1.3) and (2.11)), we obtain

$$J^\varepsilon_{0,i} + \varepsilon J^\varepsilon_{1,i} + \varepsilon^2 J^\varepsilon_{2,i} = 0, \qquad i = 1, \cdots, n, \tag{2.12}$$

where

$$J^\varepsilon_{0,i} = \int_\Omega (\tilde{p}^\varepsilon w^{i,\varepsilon} \cdot \nabla \varphi + f \cdot w^{i,\varepsilon}\varphi - \mu(\tilde{s}^\varepsilon)\, e^i \cdot \tilde{v}^\varepsilon \varphi) \, dx,$$

$$J^\varepsilon_{1,i} = -\int_\Omega \pi^{i,\varepsilon} \tilde{v}^\varepsilon \cdot \nabla(\mu(\tilde{s}^\varepsilon)\varphi) \, dx,$$

$$J^\varepsilon_{2,i} = \int_\Omega (\mu(\tilde{s}^\varepsilon)\varphi \tilde{v}^\varepsilon \cdot (\nabla w^{i,\varepsilon})\nabla \tilde{s}^\varepsilon + \mu(\tilde{s}^\varepsilon)\tilde{v}^\varepsilon \cdot (\nabla w^{i,\varepsilon})\nabla \varphi$$
$$- \mu(\tilde{s}^\varepsilon) w^{i,\varepsilon} \cdot (\nabla \tilde{v}^\varepsilon)\nabla \varphi) \, dx.$$

Using (2.6)-(2.9) and Lemmas 2.1-2.4, we find that

$$J^\varepsilon_{0,i} \to \int_\Omega (p(K\nabla \varphi)_i + (Kf)_i \varphi - \mu(s) v_i \varphi) \, dx,$$

as $\varepsilon \to 0$, and

$$J^\varepsilon_{1,i} \leqq C, \qquad J^\varepsilon_{2,i} \leqq \frac{C}{\varepsilon}.$$

Now (2.5) follows from (2.12).

3. The homogenized problem and convergence theorem. Taking into account Lemmas 2.2, 2.4, 2.5 and 2.6, we conclude that (v, p, s) is a solution to the following homogenized problem:

$$\operatorname{div} v = 0 \quad \text{in } \Omega, \tag{3.1}$$

$$v = \frac{1}{\mu(s)} K(f - \nabla p) \quad \text{in } \Omega, \tag{3.2}$$

$$v \cdot \nu = h \cdot \nu \quad \text{on } \Gamma, \tag{3.3}$$

$$\operatorname{div}(D\nabla s) - v \cdot \nabla s = 0 \quad \text{in } \Omega, \tag{3.4}$$

$$s = g \quad \text{on } \Gamma. \tag{3.5}$$

Because of (1.6), the inequality

$$0 \leqq s \leqq 1 \quad \text{a.e. in } \Omega \tag{3.6}$$

holds true.

LEMMA 3.1 ([8, p. 139], [2, p. 149]). *The matrices K and D are symmetric and positive definite.*

LEMMA 3.2 ([1, p. 231]). *Let w be the solution to the problem*

$$\operatorname{div}(w+F)=0 \quad \text{in } \Omega,$$
$$w=0 \quad \text{on } \Gamma,$$

and, for $r>1$,

$$\Phi_r(F)=\frac{\|\nabla w\|_{L^r(\Omega)}}{\|F\|_{L^r(\Omega)}}.$$

Then

$$\alpha_r=\alpha_r(\Omega)=\sup_{0\neq F\in L^r(\Omega)}\Phi_r(F)<\infty.$$

LEMMA 3.3. *Let the function μ satisfy the condition*

$$\operatorname{osc}\mu<\frac{1}{\alpha_3}\max\mu, \tag{3.7}$$

and let (v, p, s) be a solution to the problem (3.1)-(3.5). *Then $p\in W_3^1(\Omega)$, and there exists a constant $\beta>0$ (depending on Ω, $\mathcal{O}$ and μ), such that for each f and h the inequality*

$$\|\nabla p\|_{(L^3(\Omega))^n}\leqq\beta(\|f\|_{(L^3(\Omega))^n}+\|h\cdot\nu\|_{L^e(\Gamma)})$$

holds true.

Proof. The conclusion follows from Lemma 3.1 and the results of the book [1, Thm. 4.2, p. 234], where the constant β is defined.

THEOREM 3.1. *Let the function μ satisfy the condition* (3.7). *Then there exists a constant $\gamma>0$ (depending on Ω, $\mathcal{O}$, d and μ) such that under the assumption*

$$\|f\|_{(L^3(\Omega))^n}+\|h\cdot\nu\|_{L^3(\Gamma)}<\gamma, \tag{3.8}$$

problem (3.1)-(3.5) *has only one solution*:

$$(v, p, s)\in(L^2(\Omega))^n\times(L^2(\Omega)/\mathbf{R})\times(H^1(\Omega)\cap L^\infty(\Omega)).$$

Proof. Let (v^i, p^i, s^i), $i=1, 2$ be solutions and $v=v^1-v^2$, $p=p^1-p^2$, $s=s^1-s^2$. These functions satisfy the following equations and boundary conditions:

$$\operatorname{div} v=0 \quad \text{in } \Omega, \tag{3.9}$$

$$v=-\frac{1}{\mu(s_1)}K\nabla p+\frac{\mu(s_2)-\mu(s_1)}{\mu(s_1)\mu(s_2)}K(f-\nabla p_2) \quad \text{in } \Omega, \tag{3.10}$$

$$v\cdot\nu=0 \quad \text{on } \Gamma, \tag{3.11}$$

$$\operatorname{div}(D\nabla s)-v_1\cdot\nabla s-v\cdot\nabla s_2=0 \quad \text{in } \Omega, \tag{3.12}$$

$$s=0 \quad \text{on } \Gamma. \tag{3.13}$$

In the sequel, $C_k>0$ $(k=1, \cdots, 5)$ denotes a generic constant, depending generally on Ω, $\mathcal{O}$, d and μ. By use of (3.6) and Lemma 3.1, from (3.9)-(3.13) we obtain the

inequalities

$$\begin{aligned} &\|\nabla p\|_{(L^2(\Omega))^n} \leqq C_1 \|s(f-\nabla p_2)\|_{(L^2(\Omega))^n}, \\ &\|\nabla s\|_{(L^2(\Omega))^n} \leqq C_2 \|s(f-\nabla p_2)\|_{(L^2(\Omega))^n}. \end{aligned} \tag{3.14}$$

Using the inequality

$$\|s\|_{L^6(\Omega)} \leqq C_3 \|\nabla s\|_{L^2(\Omega)} \tag{3.15}$$

and Lemma 3.3, we obtain

$$\|s(f-\nabla p_2)\|_{(L^2(\Omega))^n} \leqq C_4(\|f\|_{(L^3(\Omega))^n} + \|h \cdot \nu\|_{L^3(\Gamma)})\|\nabla s\|_{(L^2(\Omega))^n}. \tag{3.16}$$

Let $s \neq 0$; from (3.15) and (3.16) we conclude that

$$1 \leqq C_5(\|f\|_{(L^3(\Omega))^n} + \|h \cdot \nu\|_{L^3(\Gamma)}).$$

Let $\gamma = 1/C_5$; assuming (3.8), we obtain a contradiction. Therefore $s = 0$ and, because of (3.14), $\nabla p = 0$; from (3.10) we obtain $v = 0$.

THEOREM 3.2. *Let μ, f and h satisfy the conditions* (3.7) *and* (3.8). *Let $(v^\varepsilon, p^\varepsilon, s^\varepsilon)$ and (v, p, s) be solutions to problem* (1.1)-(1.5) *and problem* (3.1)-(3.5), *respectively. Then there exist extensions $\tilde{v}^\varepsilon$, $\tilde{p}^\varepsilon$ and $\tilde{s}^\varepsilon$ of the functions v^ε, p^ε and s^ε, respectively, such that*

$$\begin{aligned} &\tilde{v}^\varepsilon \to v \quad \textit{weakly in } (L^2(\Omega))^n, \\ &\tilde{p}^\varepsilon \to p \quad \textit{strongly in } L^2(\Omega), \\ &\tilde{s}^\varepsilon \to s \quad \textit{strongly in } L^2(\Omega), \end{aligned}$$

as $\varepsilon \to 0$.

Proof. Because of Theorem 3.1, the functions v, p and s are unique cluster points of $\{\tilde{v}^\varepsilon\}$, $\{\tilde{p}^\varepsilon\}$ and $\{\tilde{s}^\varepsilon\}$, respectively. Therefore, the conclusion follows immediately from Lemmas 2.2 and 2.4.

Acknowledgments. The authors would like to thank the referees for their very useful comments. The first author is grateful to Professor F. Murat (University Pierre et Marie Curie, Paris) for his illuminating discussion on the subject, which indicated how to shorten the proofs and how to treat the more general case with matrix viscosity and diffusion coefficients.

REFERENCES

[1] S. N. ANTONTSEV, A. V. KAZHIKHOV AND V. MONAKHOV, *Boundary Value Problems of Mechanics of Non-Homogeneous Fluids*, Science, Novosibirsk, 1983. (In Russian.)

[2] N. S. BAKHVALOV AND A. P. PANASENKO, *Averaging of Processes in Periodic Media*, Science, Moscow, 1984. (In Russian.)

[3] G. CHAVENT, *A New Formulation of Diaphasic Incompressible Flows in Porous Media*, Lecture Notes in Mathematics, 503, Springer-Verlag, Berlin, 1976.

[4] D. CIORANESCU AND J. SAINT JEAN PAULIN, *Homogenization in open sets with holes*, J. Math. Anal. Appl., 71 (1979), pp. 590-607.

[5] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[6] L. D. LANDAU AND E. M. LIFSCHITZ, *Fluid Mechanics*, Pergamon Press, Oxford, 1968.

[7] A. MIKELIĆ AND I. AGANOVIĆ, *Homogenization in a porous medium under a nonhomogeneous boundary condition*, Bull. dell'Unione Matematica Italiana (6)6-A (1987), to appear.
[8] E. SANCHEZ-PALENCIA, *Non-Homogeneous Media and Vibration Theory*, Lecture Notes in Physics 127, Springer-Verlag, Berlin, 1980.
[9] R. TEMAM, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1979.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

004

THE NEUMANN PROBLEM FOR NONLINEAR SECOND ORDER SINGULAR PERTURBATION PROBLEMS*

BENOIT PERTHAME† AND RICHARD SANDERS‡

Abstract. Singularly perturbed second order elliptic partial differential equations with Neumann boundary conditions arise in many areas of application. These problems rarely have smooth limit solutions. In this paper, we characterize the limit solution for a wide class of such problems. We also give an abstract rate of convergence theorem and apply the abstract theorem to certain finite difference approximations.

Key words. singular perturbation, viscosity solution, viscosity inequalities

AMS(MOS) subject classifications. 35F30, 65M15

1. Introduction. In this paper, we study the singular perturbation problem for partial differential equations which have the form

$$(NP_\varepsilon)\qquad \begin{aligned} -\varepsilon\Delta u_\varepsilon + H(x, u_\varepsilon, \nabla u_\varepsilon) &= 0, \qquad x\in\Omega,\\ \frac{\partial u_\varepsilon}{\partial n}(x) &= \gamma(x), \qquad x\in\partial\Omega, \end{aligned}$$

where Ω is a bounded domain in $\mathbf{R}^d$, n is Ω's outward unit normal, u_ε is a scalar unknown and H is a continuous function on $\bar{\Omega}\times\mathbf{R}\times\mathbf{R}^d$. One application that motivates the study of singular perturbation problems of the form (NP_ε) is found in the theory of optimal stochastic control. There, H depends on the deterministic part of a stochastic ODE, a control space and a specified cost function. u_ε can be identified as the optimal cost function. The positive parameter ε in (NP_ε) can be regarded as the intensity of noise in the dynamics equation. Control problems whose trajectories reflect at a boundary give rise to Neumann problems of the type studied here; see [1] or [17] for a detailed treatment of this topic. One could ask, for instance, is the optimal cost function of a stochastic control problem related to the optimal cost of its associated deterministic problem? Are the two close in any way when the noise is small?

As $\varepsilon\downarrow 0$, it is well known that solutions of (NP_ε) do not generally converge to a classical solution of

$$(NP_0)\qquad \begin{aligned} H(x, u, \nabla u) &= 0, \qquad x\in\Omega,\\ \frac{\partial u}{\partial n}(x) &= \gamma(x), \qquad x\in\partial\Omega. \end{aligned}$$

Indeed, (NP_0) does not generally admit a C^1 solution, as can easily be seen by considering the simple example

$$\begin{aligned} \frac{du}{dx} + u &= 0, \qquad x\in[0,1],\\ \frac{du}{dx}(0) &= 0, \qquad \frac{du}{dx}(1) = 1. \end{aligned}$$

* Received by the editors April 29, 1985; accepted for publication (in revised form) March 14, 1987. This work was sponsored by the U.S. Army under contract DAAG29-80-C-0041 and by the National Science Foundation under grant MCS 82-00676.

† Centre de Mathématique Appliqué, Ecole Normale Supérieure, 45 Rue d'Ulm, 75230 Paris, France.

‡ Department of Mathematics, University of Houston, Houston, Texas 77004. This work was completed while this author was at the Ecole Normale Supérieure, 75230 Paris, France.

This example is clearly overdetermined and here the data at 0 is not compatible with the data at 1. For this reason, a more general class of solutions to (NP_0) must be sought.

A new notion of continuous weak solutions to equations of Hamilton–Jacobi type has recently been introduced. In [4] and [5], M. G. Crandall and P. L. Lions have developed techniques that have been extremely successful in establishing a number of new results concerning continuous, but not necessarily differentiable, weak solutions to first order, fully nonlinear, partial differential equations. In their work, Crandall and Lions have utilized the "vanishing viscosity method," so named because of the link to the classical technique of vanishing viscosity from fluid mechanics, and they show that the method of vanishing viscosity gives rise to a specific notion of a "viscosity" weak solution.

In [15], and here as well, the notion of a viscosity solution for the generally overdetermined Neumann problem (NP_0) is given and is shown to include all L^∞ ε-limits of solutions to (NP_ε). All L^∞ ε-limits are shown to satisfy the so-called viscosity inequalities of § 2; additional details in this direction can be found in [15]. Remarkably (with additional hypotheses of course), these viscosity inequalities uniquely determine all such limits. In § 3, we introduce what we call "approximate viscosity solutions" and we show there that any reasonable approximate viscosity solution is approximately equal to the viscosity solution of (NP_0). More precisely, we give an abstract minimal rate-of-convergence theorem, Theorem 2, for approximate viscosity solutions to (NP_0). We also show that this rate is essentially sharp. A particular application of Theorem 2 gives an easy to determine measure of how far the solution of (NP_ε) can be away from the viscosity solution of (NP_0). In § 4, the abstract rate-of-convergence theorem of § 3 is applied to numerical approximations which are obtained from a class of finite difference schemes. Moreover, we show in § 4 that these schemes have "computable" solutions and we motivate how they can be obtained.

The reader is encouraged to see [2], [18] and [19] where similar results as those above are obtained for divergence form singular perturbation problems with mixed or Dirichlet boundary conditions. See also [6], [7] and [21] for a further treatment of approximations for time-dependent Hamilton–Jacobi equations without spatial boundaries.

2. Viscosity solutions. As mentioned in the previous section, as $\varepsilon \downarrow 0$, the corresponding solutions to (NP_ε) do not in general converge to a classical C^1 solution of (NP_0). In this section we offer a characterization of viscous limits to (NP_0) and we show that this characterization often allows for only one solution in the class of continuous functions. Throughout, we shall assume that Ω is a bounded domain in $\mathbf{R}^d$ which has a C^2 boundary $\partial\Omega$. The *outward* normal of Ω at a point $x \in \partial\Omega$ will be denoted by $n(x)$ and we write the outward normal derivative of φ at $x \in \partial\Omega$ as $(\partial\varphi/\partial n)(x)$.

We should like to mention that previous to the writing of this paper P. L. Lions had introduced the same viscosity characterization of solutions to (NP_0) as we give below (see [15]). For this reason, we borrow much of the notation and hypotheses of [15] and in this section we omit all proofs but those which motivate the results of the next sections.

We now state the viscosity characterization (see Proposition 1) of continuous weak solutions to (NP_0).

DEFINITION 1. Suppose that $H(x, u, p) \in C(\bar{\Omega} \times \mathbf{R} \times \mathbf{R}^d)$ and $u(x) \in C(\bar{\Omega})$. We say that:

(a) $u(x)$ is a *viscosity subsolution* of (NP_0) if for all test functions $\varphi \in C^1(\mathbf{R}^d)$

with $(\partial\varphi/\partial n)(x) \geqq \gamma(x)$, we have

$$H(x_0, u(x_0), \nabla\varphi(x_0)) \leqq 0,$$

where $x_0 \in \bar{\Omega}$ satisfies

$$u(x_0) - \varphi(x_0) = \max_{x\in\bar{\Omega}} (u(x) - \varphi(x)).$$

(b) $u(x)$ is a *viscosity supersolution* of (NP_0) if for all test functions $\varphi \in C^1(\mathbf{R}^d)$ with $(\partial\varphi/\partial n)(x) \leqq \gamma(x)$, we have

$$H(x_0, u(x_0), \nabla\varphi(x_0)) \geqq 0,$$

where $x_0 \in \bar{\Omega}$ satisfies

$$u(x_0) - \varphi(x_0) = \min_{x\in\bar{\Omega}} (u(x) - \varphi(x)).$$

(c) $u(x)$ is a *viscosity solution* of (NP_0) if it satisfies both (a) and (b) above.

The fact that our test functions are required to satisfy $(\partial\varphi/\partial n)(x) \geqq \gamma(x)$ (resp. $(\partial\varphi/\partial n)(x) \leqq \gamma(x)$), in our definition of viscosity subsolution (resp. supersolution), may at first seem superfluous. This is, however, precisely the mechanism that "sees" the Neumann boundary conditions when vanishing viscosity is taken into account (see Proposition 1 below).

Remark 2.1. Any C^1 solution of (NP_0) is also a viscosity solution. This fact is nontrivial only for the case when $\max(u-\varphi)$ or $\min(u-\varphi)$ is attained for some $x_0 \in \partial\Omega$. To see that u must indeed be a viscosity subsolution, take an arbitrary $\varphi \in C^1(\mathbf{R}^d)$ with $(\partial\varphi/\partial n)(x) \geqq \gamma(x)$. First choose a sequence $\{\varphi_m\}$, such that for every m, $\varphi_m(x_0) = \varphi(x_0)$, $\varphi_m(x) > \varphi(x)$ for $x \neq x_0$, $(\partial\varphi_m/\partial n)(x) \geqq (\partial\varphi/\partial n)(x)$ and with $\varphi_m \to \varphi$ in C^1 as $m \to \infty$. We then have that $(u-\varphi)(x_0) = \max(u-\varphi_m)$ and x_0 is the point where the *strict* maximum of $u-\varphi_m$ is attained. Next, for any fixed m, choose a sequence $\{\varphi_m^n\}$ such that $(\partial\varphi_m^n/\partial n)(x) > (\partial\varphi_m/\partial n)(x)$ and with $\varphi_m^n \to \varphi_m$ in C^1 as $n \to \infty$. Denoting by x_n the points where $\max(u-\varphi_m^n)$ is attained, we must have that $x_n \to x_0$ as $n \to \infty$. This is true because $(u-\varphi_m)(x_0)$ is a strict maximum of $u-\varphi_m$. For $x \in \partial\Omega$, we also have that $(\partial/\partial n)(u-\varphi_m^n)(x) < 0$, which implies $x_n \in \Omega^0$. Therefore, since now x_n is an interior maximum of $u-\varphi_m^n$, $\nabla u(x_n) = \nabla\varphi_m^n(x_n)$, and so by taking limits we have

$$H(x_0, u(x_0), \nabla\varphi(x_0)) = \lim_{m,n} H(x_n, u(x_n), \nabla\varphi_m^n(x_n)) = 0.$$

Remark 2.2. Obviously, the converse of Remark 2.1 is false. That is, a smooth viscosity solution need not satisfy the boundary conditions of (NP_0).

PROPOSITION 1. *Let $u_\varepsilon \in C^2(\bar{\Omega})$ be a solution of (NP_ε) and suppose that $\varphi \in C^2(\bar{\Omega})$. Then*:

(a) *For $(\partial\varphi/\partial n)(x) \geqq \gamma(x)$ and $u_\varepsilon(x_0) - \varphi(x_0) = \max_{x\in\bar{\Omega}}(u_\varepsilon(x) - \varphi(x))$ we have that $H(x_0, u_\varepsilon(x_0), \nabla\varphi(x_0)) \leqq \varepsilon\Delta\varphi(x_0)$.*

(b) *For $(\partial\varphi/\partial n)(x) \leqq \gamma(x)$ and $u_\varepsilon(x_0) - \varphi(x_0) = \min_{x\in\bar{\Omega}}(u_\varepsilon(x) - \varphi(x))$ we have that $H(x_0, u_\varepsilon(x_0), \nabla\varphi(x_0)) \geqq \varepsilon\Delta\varphi(x_0)$.*

If in addition, we have that $u_\varepsilon \to u$ in $L^\infty(\bar{\Omega})$ for some sequence $\varepsilon \downarrow 0$, then:

(c) *$u = \lim_\varepsilon u_\varepsilon$ is a viscosity solution, that is, u satisfies Definition* 1(c).

The proof of Proposition 1 can be found in [15]; however, the interested reader can easily reproduce its proof by taking limits as in Remark 2.1.

Before stating the main result of this section, we give a simple lemma.

LEMMA 1. *Let Ω be a bounded domain in $\mathbf{R}^d$ having a C^2 boundary $\partial\Omega$. Then*
(a) *There exists a constant $C_\Omega < \infty$ such that for all $x \in \partial\Omega$,*

$$C_\Omega \geqq \sup_{y \in \bar{\Omega}} \left(\frac{-(x-y)\cdot n(x)}{|x-y|^2} \right).$$

(b) *There exists a function $w \in C^2(\bar{\Omega})$ such that*

$$\frac{\partial w}{\partial n}(x) = \max(C_\Omega, 0), \quad x \in \partial\Omega,$$

$$|\nabla w(x)| \leqq \max(C_\Omega, 0), \quad x \in \bar{\Omega}.$$

Proof. $C_\Omega = w(x) \equiv 0$ would suffice in the case of convex Ω. For nonconvex Ω, (a) is shown in [12]. (b) can be shown by constructing a particular example. Under the hypotheses of the lemma, it is known that the distance function $d(x) = d(x; \partial\Omega)$ is C^2 in a neighborhood of $\partial\Omega$ [23], [10]. That is, $d(x) \in C^2(\Omega_\tau)$, where $\Omega_\tau = \{x \in \bar{\Omega}: d(x) < \tau\}$ and $\tau > 0$ is chosen sufficiently small. Set $0 < \tau_0 < \tau$ and verify that

$$w(x) = \begin{cases} \dfrac{C_\Omega}{3\tau_0^2}(\tau_0 - d(x))^3 & \text{if } x \in \bar{\Omega}_{\tau_0}, \\ 0 & \text{if } x \in \bar{\Omega}\Omega_{\tau_0}, \end{cases}$$

is a particular example that satisfies (b).

Now, consider the following set of assumptions.

Assumption A. $H(x, u, p)$ is strictly increasing in u for all $x \in \bar{\Omega}$ and uniformly for $p \in \mathbf{R}^d$. That is, for all $R > 0$ and $-R \leqq v \leqq u \leqq R$, there exists a $\mu_R > 0$ such that

$$H(x, u, p) - H(x, v, p) \geqq \mu_R(u - v).$$

Assumption B. Let $\alpha, \beta \in \mathbf{R}^d$ satisfy $|\alpha|, |\beta| \leqq \max(C_\Omega, 0)$, where C_Ω is as defined in the previous lemma. Then, for all such α, β and all $x \in \bar{\Omega}$, $x + \xi \in \bar{\Omega}$, all $|u| \leqq R$ and any $\lambda > 1$, assume that

$$\left| H\left(x + \xi, u, \lambda\xi + \frac{\lambda}{2}|\xi|^2\alpha + O(\xi)\right) - H\left(x, u, \lambda\xi + \frac{\lambda}{2}|\xi|^2\beta\right) \right| \leqq \omega_R(\lambda|\xi|^2 + |\xi|),$$

where $\omega_R(s)$ is some function such that $\lim_{s \downarrow 0} \omega_R(s) = 0$.

Remark 2.3. Assumptions A and B are standard (see [5], [8] and [15]). In the following theorem, Assumption B may always be relaxed so that $\alpha = \beta \equiv 0$ and $0(\xi) \equiv 0$ except for x in a neighborhood of $\partial\Omega$. Assuming additional regularity on the class of solutions allows Assumption B to be neglected entirely.

THEOREM 1. *Suppose that $H(x, u, p) \in C(\bar{\Omega} \times \mathbf{R} \times \mathbf{R}^d)$ and that it satisfies Assumption* A *above. Let $u \in C(\bar{\Omega})$ be a viscosity subsolution of (NP_0) and let $v \in C(\bar{\Omega})$ be a viscosity supersolution of (NP_0). Finally, assume one from the following three sets of hypotheses*:

(i) *Ω is convex and Assumption* B *holds with $\alpha = \beta \equiv 0$.*

(ii) *Assumption* B *is satisfied.*

(iii) *Either u or v is Lipschitz continuous.*

We then have that

$$\max_{x \in \bar{\Omega}} (u(x) - v(x)) \leqq 0.$$

Obviously, establishing this result would imply that a viscosity *solution* to (NP_0) is unique in the specified class of functions.

Proof. Given a $\delta > 0$, define the function $\phi^\delta(x, y)$ by

$$\phi^\delta(x, y) = \rho(x)\rho(y)|x-y|^2/\delta, \tag{2.1}$$

where $\rho(x) = \exp(w(x))$ and $w(x)$ satisfies the second conclusion of Lemma 1. For $x \in \partial\Omega$ and any fixed $y_0 \in \bar{\Omega}$, observe that

$$\frac{\partial}{\partial n_x}\phi^\delta(x, y_0) = \phi^\delta(x, y_0)\left\{\frac{\partial w}{\partial n}(x) + \frac{2(x-y_0)\cdot n}{|x-y_0|^2}\right\},$$

and Lemma 1 implies that the bracketed term above is nonnegative. Therefore,

$$\frac{\partial}{\partial n_x}\phi^\delta(x, y_0) \geqq 0$$

and, similarly, for $y \in \partial\Omega$ and any fixed $x_0 \in \bar{\Omega}$

$$\frac{\partial}{\partial n_y}\phi^\delta(x_0, y) \geqq 0.$$

Now, choose $\psi \in C^2(\bar{\Omega})$ such that $(\partial\psi/\partial n)(x) = \gamma(x)$. By the construction above, we have that for any fixed $y_0 \in \bar{\Omega}$

$$\phi_1(x) = \phi^\delta(x, y_0) + \psi(x)$$

is an admissible test function according to Definition 1(a) and similarly for fixed $x_0 \in \bar{\Omega}$

$$\phi_2(y) = -\phi^\delta(x_0, y) + \psi(y)$$

is admissible according to Definition 1(b).

The next step is to note the obvious inequality

$$\max_{x\in\bar{\Omega}}(u(x)-v(x)) \leqq \max_{\substack{x\in\bar{\Omega}\\ y\in\bar{\Omega}}}(u(x)-v(y)-(\phi^\delta(x, y)+\psi(x)-\psi(y))). \tag{2.2}$$

We denote by x_δ, y_δ the points in $\bar{\Omega}$ where the right-hand side of (2.2) is attained and we rewrite (2.2) as

$$\max_{x\in\bar{\Omega}}(u(x)-v(x)) \leqq u(x_\delta)-v(y_\delta)-(\phi^\delta(x_\delta, y_\delta)+\psi(x_\delta)-\psi(y_\delta)). \tag{2.3}$$

Using (2.3), we easily arrive at

$$\begin{aligned}\phi^\delta(x_\delta, y_\delta) &\leqq |u(x_\delta)-u(y_\delta)| + |\psi(x_\delta)-\psi(y_\delta)|,\\ \phi^\delta(x_\delta, y_\delta) &\leqq |v(x_\delta)-v(y_\delta)| + |\psi(x_\delta)-\psi(y_\delta)|,\end{aligned} \tag{2.4}$$

and recalling the definition of $\phi^\delta(x, y)$, (2.4) gives us that

$$|x_\delta - y_\delta| \leqq \text{const.}\sqrt{\delta}. \tag{2.5}$$

Furthermore, since u, (or v), and ψ are continuous, (2.4) combined with (2.5) shows that

$$\lim_{\delta\downarrow 0}\phi^\delta(x_\delta, y_\delta) = 0. \tag{2.6}$$

The object now is to show that the right-hand side of (2.2) can be made arbitrarily small by choosing δ sufficiently small. From above, we see that the test functions defined as

$$\begin{aligned}\phi_1(x) &= v(y_\delta) - \psi(y_\delta) + \phi^\delta(x, y_\delta) + \psi(x),\\ \phi_2(y) &= u(x_\delta) - \psi(x_\delta) - \phi^\delta(x_\delta, y) + \psi(y),\end{aligned}$$

are admissible according to Definition 1(a) and Definition 1(b), respectively. Inserting these into Definition 1, and using the fact that u is a viscosity subsolution and v is a viscosity supersolution, allows us to conclude that

$$H(x_\delta, u(x_\delta), \nabla_x \phi_1(x_\delta)) \leqq 0$$

and

$$H(y_\delta, v(y_\delta), \nabla_y \phi_2(y_\delta)) \geqq 0$$

because x_δ satisfies

$$u(x_\delta) - \phi_1(x_\delta) = \max_{x \in \bar{\Omega}} (u(x) - \phi_1(x))$$

and y_δ satisfies

$$v(y_\delta) - \phi_2(y_\delta) = \min_{y \in \bar{\Omega}} (v(y) - \phi_2(y)).$$

Combining the inequalities above and rearranging, we obtain

$$\begin{aligned} &H(x_\delta, u(x_\delta), \nabla_x \phi_1(x_\delta)) - H(x_\delta, v(y_\delta), \nabla_x \phi_1(x_\delta)) \\ &\quad \leqq H(y_\delta, v(y_\delta), \nabla_y \phi_2(y_\delta)) - H(x_\delta, v(y_\delta), \nabla_x \phi_1(x_\delta)). \end{aligned} \tag{2.7}$$

By a direct calculation, the right-hand side of (2.7) can be written as

$$\begin{aligned} &H\left(y_\delta, v(y_\delta), \lambda(x_\delta - y_\delta) + \frac{\lambda}{2}|x_\delta - y_\delta|^2 \beta + \nabla \psi(y_\delta)\right) \\ &\quad - H\left(x_\delta, v(y_\delta), \lambda(x_\delta - y_\delta) + \frac{\lambda}{2}|x_\delta - y_\delta|^2 \alpha + \nabla \psi(x_\delta)\right), \end{aligned} \tag{2.8}$$

where

$$\lambda = 2\rho(x_\delta)\rho(y_\delta)/\delta, \quad \alpha = \nabla w(x_\delta), \quad \beta = \nabla w(y_\delta).$$

(Recall from Lemma 1 that if Ω is convex we may assume that $\alpha = \beta \equiv 0$ and $\rho(x) = \rho(y) \equiv 1$.) Assumption B allows (2.8) to be bounded above by

$$\omega_R(\lambda |x_\delta - y_\delta|^2 + |x_\delta - y_\delta|), \tag{2.9}$$

where $R = \max(|u|_\infty, |v|_\infty)$.

To complete the proof of conclusions (i) and (ii), we again use inequality (2.3) to write

$$\max_{x \in \bar{\Omega}} (u(x) - v(x)) \leqq u(x_\delta) - v(y_\delta) + |\psi(x_\delta) - \psi(y_\delta)|,$$

which is bounded above by

$$\max((u(x_\delta) - v(y_\delta)), 0) + |\psi(x_\delta) - \psi(y_\delta)|. \tag{2.10}$$

Assumption A applied to the left-hand side of (2.7) combined with (2.8) and (2.9), allows us to bound (2.10) by

$$\frac{1}{\mu_R} \omega_R(\lambda |x_\delta - y_\delta|^2 + |x_\delta - y_\delta|) + |\psi(x_\delta) - \psi(y_\delta)|. \tag{2.11}$$

Recalling that $\lambda |x_\delta - y_\delta|^2 = 2\phi^\delta(x_\delta, y_\delta)$, (2.6) along with (2.5) show that (2.11) tends to zero as δ tends to zero, thereby proving that

$$\max_{x \in \bar{\Omega}} (u(x) - v(x)) \leqq 0.$$

To establish (iii), observe that if u (or v) is Lipschitz continuous, inequality (2.4) leads to an improvement of (2.5). That is, we may conclude that

$$|x_\delta - y_\delta| \leqq \text{const. } \delta. \tag{2.12}$$

This improved estimate implies that the p term of $H(\cdot, \cdot, p)$ in (2.8) remains bounded. Therefore, conclusion (iii) follows by noting the uniform continuity of $H(x, u, p)$ on a compact subset of $\bar{\Omega} \times \mathbf{R} \times \mathbf{R}^d$.

Remark 2.4. The combined results of Proposition 1 and Theorem 1 imply that *if* the family $\{u_\varepsilon\}_{\varepsilon>0}$ of solutions to (NP_ε) is relatively compact in L^∞, then $\lim_{\varepsilon \downarrow 0} u_\varepsilon$ exists in L^∞. For further results concerning the compactness of $\{u_\varepsilon\}_{\varepsilon>0}$ see [13] or [14].

3. Viscosity approximations and a rate of convergence. In this section we consider the rate at which certain approximations converge to the viscosity solution of (NP_0). We show in a precise sense below that if an approximation "almost" satisfies the viscosity inequalities of Definition 1, then the approximation is "almost" equal to its associated viscosity limit solution. The abstract rate of convergence theorem given in this section is then applied in § 4 to particular approximations generated by a class of numerical schemes.

Before making a precise statement of "almost satisfies the viscosity inequalities," recall the definitions of the test functions used in the proof of Theorem 1:

$$\phi^\delta(x, y) = \rho(x)\rho(y)|x - y|^2/\delta, \tag{3.1}$$

where $\delta > 0$, $\rho(x) = \exp(w(x))$ and $w(x)$ satisfies the second conclusion of Lemma 1. Also recall the function ψ, which satisfies

$$\begin{aligned} &\psi(x) \in C^2(\bar{\Omega}), \\ &\Delta\psi = 0 \quad \text{in } \Omega, \\ &\frac{\partial \psi}{\partial n}(x) = \gamma(x) \quad \text{on } \partial\Omega, \end{aligned} \tag{3.2}$$

and the specific test functions

$$\phi_1(x) = \phi^\delta(x, y_0) + \psi(x), \tag{3.3a}$$

$$\phi_2(y) = -\phi^\delta(x_0, y) + \psi(y), \tag{3.3b}$$

where x_0, y_0 are arbitrary fixed points in $\bar{\Omega}$. Notice that $\nabla\psi$ is uniquely determined by (3.2).

We now give the following definition.

DEFINITION 2. Suppose that $H(x, u, p) \in C(\bar{\Omega} \times \mathbf{R} \times \mathbf{R}^d)$ and $u_\varepsilon \in C(\bar{\Omega})$. We say that:

(a) u_ε is an *approximate viscosity subsolution of order* ε to (NP_0) if there exists a family of test functions of the form (3.3a) so that

$$H(x_0, u_\varepsilon(x_0), \nabla_x \phi_1(x_0)) \leqq \varepsilon \Delta_x \phi_1(x_0) + \underline{C}\varepsilon,$$

where $x_0 \in \bar{\Omega}$ satisfies

$$u_\varepsilon(x_0) - \phi_1(x_0) = \max_{x \in \bar{\Omega}} (u_\varepsilon(x) - \phi_1(x))$$

and $\underline{C}$ is some fixed constant.

(b) u_ε is an *approximate viscosity supersolution of order* ε to (NP_0) if there exists a family of test functions of the form (3.3b) so that

$$H(y_0, u_\varepsilon(y_0), \nabla_y \phi_2(y_0)) \geqq \varepsilon \Delta_y \phi_2(y_0) - \underline{C}\varepsilon,$$

where $y_0 \in \bar{\Omega}$ satisfies

$$u_\varepsilon(y_0) - \phi_2(y_0) = \min_{x \in \bar{\Omega}} (u_\varepsilon(x) - \phi_2(x)).$$

(c) u_ε is an *approximate viscosity solution of order* ε if it satisfies both (a) and (b) above.

In the proof of Theorem 1 we showed that $(\partial \phi_1 / \partial n_x)(x) \geqq \gamma(x)$ and $(\partial \phi_2 / \partial n_y)(y) \leqq \gamma(y)$; therefore, the statement of Proposition 1 implies that if u_ε is a C^2 solution of (NP_ε), then it is also an approximate viscosity solution of (NP_0) as defined above.

With Definition 2, we now state the following.

THEOREM 2. *In addition to Assumption* A *of Theorem* 1, *assume for ease of presentation that* $\gamma(x) \equiv 0$. *Furthermore, assume that* (NP_0) *admits a Lipschitz continuous viscosity solution* u, *with say Lipschitz constant* L, *and assume that* $H(x, u, p)$ *is locally Lipschitz continuous. Then, for any approximate viscosity solution to* (NP_0), *say* u_ε, *we have that*

$$|u_\varepsilon - u|_\infty \leqq \frac{1}{\mu_{R_0}} [(8dL_H L(1 + 2C_\Omega L)\varepsilon)^{1/2} + O(\varepsilon)]$$

where

$$R_0 = \sup_{1 \geqq \varepsilon > 0} |u_\varepsilon|_\infty,$$

$$L_H = \sup_{\substack{x_1, x_2 \in \bar{\Omega} \\ |u| \leqq R_0 \\ |P_1|, |P_2| \leqq 3L}} \left[\frac{|H(x_1, u, P_1) - H(x_2, u, P_2)|}{|x_1 - x_2| + |P_1 - P_2|} \right].$$

The definition of C_Ω *is given in Lemma* 1.

Remark 3.1. We may replace the assumption that (NP_0) admits a Lipschitz continuous viscosity solution by the assumption that u_ε is Lipschitz continuous, uniformly in $\varepsilon > 0$.

Remark 3.2. The interested reader can easily modify the following proof to include inhomogeneous boundary data to obtain the same $\sqrt{\varepsilon}$ rate of convergence. Relaxing the hypothesis on $H(x, u, p)$ and the regularity of u can also be done to obtain a more general (and slower) rate of convergence. This, however, will not be done here.

Proof of Theorem 2. Mimicking the proof of Theorem 1, we arrive at the analogue of inequality (2.7):

$$\begin{aligned} &H(x_\delta, u_\varepsilon(x_\delta), \nabla_x \phi_1(x_\delta)) - H(x_\delta, u(y_\delta), \nabla_x \phi_1(x_\delta)) \\ &\quad \leqq H(y_\delta, u(y_\delta), \nabla_y \phi_2(y_\delta)) - H(x_\delta, u(y_\delta), \nabla_x \phi_1(x_\delta)) + \varepsilon \Delta_x \phi_1(x_\delta) + \underline{C}\varepsilon, \end{aligned} \tag{3.4}$$

where $\nabla_x \phi_1$ and $\nabla_y \phi_2$ are given by (3.3) (with $\psi \equiv 0$) and x_δ and y_δ are as in (2.3). Recalling (2.4) and using the fact that u (or u_ε) is Lipschitz continuous, we have that

$$\phi^\delta(x_\delta, y_\delta) \leqq |u(x_\delta) - u(y_\delta)| \leqq L|x_\delta - y_\delta|, \tag{3.5}$$

or

$$(\phi^\delta(x_\delta, y_\delta) \leqq |u_\varepsilon(x_\delta) - u_\varepsilon(y_\delta)| \leqq L|x_\delta - y_\delta|),$$

which, with the definition of ϕ^δ, gives us that

$$|x_\delta - y_\delta| \leqq \frac{L}{\rho(x_\delta)\rho(y_\delta)} \delta. \tag{3.6}$$

Furthermore, a direct calculation shows that

$$|\nabla_x \phi_1(x) - \nabla_y \phi_2(y)| \leqq 2C_\Omega \phi^\delta(x, y),$$

where C_Ω is as in Lemma 1. This inequality, along with (3.5) shows that

$$|\nabla_x \phi_1(x_\delta) - \nabla_y \phi_2(y_\delta)| \leqq 2C_\Omega L|x_\delta - y_\delta|. \tag{3.7}$$

Returning now to inequality (3.4), we use (3.7), Assumption A and the fact that H is (locally) Lipschitz continuous to obtain

$$\mu_{R_0} \max((u_\varepsilon(x_\delta) - u(y_\delta)), 0) \leqq L_H(1 + 2C_\Omega L) \cdot |x_\delta - y_\delta| + \varepsilon \Delta_x \phi_1(x_\delta) + \underline{C}\varepsilon.$$

Calculating $\Delta_x \phi_1$ and inserting (3.6) into the right-hand side above, we find that

$$\begin{aligned} &\mu_{R_0} \max((u_\varepsilon(x_\delta) - u(y_\delta)), 0) \\ &\leqq \left[2d\varepsilon \left(\frac{\rho(x_\delta)\rho(y_\delta)}{\delta} \right) + \hat{L}L \left(\frac{\delta}{\rho(x_\delta)\rho(y_\delta)} \right) \right] + \text{const.}\,(\varepsilon + \varepsilon\delta), \end{aligned} \tag{3.8}$$

where $\hat{L} = L_H(1 + 2C_\Omega L)$. By setting

$$\frac{\delta}{\rho(x_\delta)\rho(y_\delta)} = \left(\frac{2d}{\hat{L}L} \varepsilon \right)^{1/2},$$

which can be done for $\hat{L}L \neq 0$ by the continuity of the left-hand side with respect to δ, we minimize the bracketed term in (3.8). This yields

$$\mu_{R_0} \max((u_\varepsilon(x_\delta) - u(y_\delta)), 0) \leqq (8d\hat{L}L\varepsilon)^{1/2} + \text{const.}\,(\varepsilon + \varepsilon^{3/2}),$$

and using the fact that

$$\max_{x \in \bar{\Omega}} (u_\varepsilon(x) - u(x)) \leqq u_\varepsilon(x_\delta) - u(y_\delta),$$

as done in the proof of Theorem 1, we have established the desired result for $\max(u_\varepsilon - u)$.

An identical estimate can be obtained for $\max(u - u_\varepsilon)$ by a similar argument and so the proof of Theorem 2 is complete.

Remark 3.3. When the domain Ω is convex and $\psi \equiv 0$, the term $O(\varepsilon)$ in the estimate of Theorem 2 is precisely $\underline{C}\varepsilon$. In addition, if the approximate viscosity solution is the solution of (NP_ε), the constant $\underline{C}$ is zero.

Next, we show that the order of the rate of convergence obtained above cannot in general be improved. To see this, consider the example

$$\begin{aligned} &-\varepsilon \frac{d^2 u_\varepsilon}{dx^2} + u_\varepsilon = 0, \\ &\frac{du_\varepsilon}{dx}(0) = 0, \qquad \frac{du_\varepsilon}{dx}(1) = 1. \end{aligned} \tag{3.9}$$

The exact solution of (3.9) is given by

$$u_\varepsilon(x) = \sqrt{\varepsilon}\, \cosh(x/\sqrt{\varepsilon}) / \sinh(1/\sqrt{\varepsilon})$$

and it is an easy exercise to show that $u_\varepsilon \to 0$ uniformly as $\varepsilon \downarrow 0$. In fact, one easily finds that

$$|u_\varepsilon - 0|_\infty = \sqrt{\varepsilon}\{1 + O(\exp(-2/\sqrt{\varepsilon}))\},$$

which is exactly the order obtained by Theorem 2. We should mention, however, that the rate constant of Theorem 2 is not the best possible.

We conclude this section by analyzing the specific example:

$$
\begin{aligned}
&-\varepsilon\frac{d^2u_\varepsilon}{dx^2}+\left(\frac{1}{2}\frac{du_\varepsilon}{dx}\right)^2+u_\varepsilon=0,\\
&\frac{du_\varepsilon}{dx}(0)=\gamma_0, \qquad \frac{du_\varepsilon}{dx}(1)=-\gamma_0.
\end{aligned}
\tag{3.10}
$$

Setting $\varepsilon=0$ and solving the reduced differential equation, we find that $u=\lim_\varepsilon u_\varepsilon$ should be built from functions having the form $-(x-c)^2$ and 0. The objective now is to piece things together in such a way that the constructed function satisfies the viscosity inequalities of Definition 1.

We have three basic cases (which depend on γ_0). Set

$$
u_L(x)=-\left(x-\frac{\gamma_0}{2}\right)^2, \qquad u_R(x)=-\left(x+\frac{\gamma_0}{2}-1\right)^2,
$$

and note that u_L satisfies the left boundary condition of (3.10) and u_R satisfies the right boundary conditions.

Case 1. For $1\geqq\gamma_0\geqq 0$, consider the candidate limit solution:

$$
u_1(x)=\begin{cases} u_L(x), & 0\leqq x\leqq\dfrac{\gamma_0}{2},\\[2ex] 0, & \dfrac{\gamma_0}{2}\leqq x\leqq 1-\dfrac{\gamma_0}{2},\\[2ex] u_R(x), & 1-\dfrac{\gamma_0}{2}\leqq x\leqq 1.\end{cases}
$$

The analysis of this case is trivial since $u_1(x)$ is a classical C^1 solution to the reduced problem. By Remark 2.1, it must therefore be a viscosity solution.

Case 2. For $\gamma_0>1$, consider the candidate limit solution:

$$
u_2(x)=\begin{cases} u_L(x), & 0\leqq x\leqq\frac{1}{2},\\ u_R(x), & \frac{1}{2}\leqq x\leqq 1.\end{cases}
$$

Obviously, we need only check the viscosity inequalities at $x_0=\frac{1}{2}$, the corner of u_2. In this case, however, $\min(u_2-\phi)$ cannot occur at $x_0=\frac{1}{2}$ for any C^1 function ϕ. If $\max(u_2-\phi)$ occurs at $x_0=\frac{1}{2}$, it is easy to check that we must have $(u_L)_x(\frac{1}{2})\geqq\phi_x(\frac{1}{2})\geqq(u_R)_x(\frac{1}{2})$. Computing these derivatives, we have that all possible values of $\phi_x(\frac{1}{2})$ lie in the interval $[1-\gamma_0,\ \gamma_0-1]$, in which case

$$
(\tfrac{1}{2}\phi_x(\tfrac{1}{2}))^2+u_2(\tfrac{1}{2})=(\tfrac{1}{2}\phi_x(\tfrac{1}{2}))^2-(\tfrac{1}{2}(1-\gamma_0))^2\leqq 0.
$$

Therefore, u_2 is a viscosity solution.

Case 3. For $\gamma_0<0$, consider the candidate:

$$
u_3(x)=0.
$$

Here, u_3 does not take on its boundary condition at $x=0$ or at $x=1$. However, $\max(u_3-\phi)$ cannot occur at $x=0$ for any admissible test function, $(\partial\phi/\partial n(0)\geqq-\gamma_0)$. If on the other hand, $\min(u_\varepsilon-\phi)$ is attained at $x_0=0$, we must have that $\phi_x(0)$ lies in $[\gamma_0, 0]$ and in this case

$$
(\tfrac{1}{2}\phi_x(0))^2+u_3(0)\geqq 0.
$$

A similar argument shows that u_3 satisfies the viscosity inequalities if $\max(u_3-\phi)$ is attained at $x_0=1$.

In these specific examples, we have demonstrated that these candidate limit solutions are viscosity solutions of (3.10) since they satisfy Definition 1(c). They are furthermore Lipschitz continuous and so by Proposition 1 and Remark 3.2 of Theorem 2, they satisfy $|u_\varepsilon - u|_\infty \leqq \text{const.}\sqrt{\varepsilon}$ where u_ε is the exact solution of problem (3.10). However, for these examples (as well as other nonlinear examples) there is evidence that indicates a convergence rate faster than the $\sqrt{\varepsilon}$, [3]. We believe that there is a yet undiscovered mechanism that links certain nonlinearities in H to diffusion which often gives rise to a faster rate of convergence than Theorem 2 predicts.

4. Numerical approximations. In this section, we introduce and analyze a class of numerical schemes that generate approximations of the viscosity *limit* solution to the one-dimensional version of (NP_ε), which we write here as:

$$\begin{aligned} &-\varepsilon\frac{d^2u_\varepsilon}{dx^2}+H\left(x, u_\varepsilon, \frac{du_\varepsilon}{dx}\right)=0, \\ &-\frac{du_\varepsilon}{dx}(0)=\gamma_0, \qquad \frac{du_\varepsilon}{dx}(1)=\gamma_1. \end{aligned} \tag{4.1}$$

Throughout this section, we make the following assumptions concerning $H(x, u, p)$, which for ease of presentation only, is assumed C^1 smooth.

Assumption A′. For all $x\in[0,1]$, $|u|=R$ and $|P|\leqq K$, there exists a $\mu_K>0$ and an $0\leqq\eta_1<1$, such that

$$\frac{\partial}{\partial u}H(x, u, p)\geqq \mu_K/(\max(R,1))^{\eta_1}.$$

Assumption B′. For all $x\in[0,1]$ and $|p|=K$, there exists an $0\leqq\eta_2<1$ and a constant $C(|u|)$ such that

$$\left|\frac{\partial}{\partial x}H(x, u, p)\right| \leqq \mu_K(\max(K,1))^{\eta_2}C(|u|).$$

Assumption A′ is merely a refined version of Assumption A of § 2. Assumption B′ guarantees that the viscosity limit solution of (4.1) is Lipschitz continuous and therefore supercedes Assumption B of § 2.

The numerical approximations that are considered here are built from a piecewise linear interpolation of grid values $\{u_j\}_{j=0}^J$. That is, we partition the interval $[0,1]$ as $\bigcup_{j=0}^{J-1}[x_j, x_{j+1}]$, where we shall assume that

$$2(x_j-x_{j-1})\geqq(x_{j+1}-x_j)\geqq\tfrac{1}{2}(x_j-x_{j-1}),$$

and then define $u^\Delta(x)$ by

$$u^\Delta(x)=\sum_{j=0}^{J}u_jT_j(x), \tag{4.2}$$

where

$$T_j(x)=\begin{cases}(x-x_{j-1})/(x_j-x_{j-1}) & \text{if } x\in[x_{j-1}, x_j],\\ (x_{j+1}-x)/(x_{j+1}-x_j) & \text{if } x\in[x_j, x_{j+1}],\\ 0 & \text{otherwise.}\end{cases}$$

In (4.2) the superscript Δ is to represent a measure of grid refinement and we set it equal to $\max_{0\leqq j\leqq J-1}(x_{j+1}-x_j)$. For each $0\leqq j\leqq J$, the grid values u_j are required to

satisfy the difference scheme

$$\begin{aligned} &\bar{H}(x_j, u_j, D^+u_j, D^-u_j) = 0, \\ &-D^-u_0 = \gamma_0, \qquad D^+u_J = \gamma_1, \end{aligned} \tag{4.3}$$

where $D^+u_j = (u_{j+1} - u_j)/(x_{j+1} - x_j)$, $D^-u_j = (u_j - u_{j-1})/(x_j - x_{j-1})$, and $\bar{H}(x, u, p_1, p_2)$ is some difference operator that does not explicitly depend on any grid parameter. $\bar{H}(x, u, p_1, p_2)$ is assumed to be locally Lipschitz continuous and it is also assumed to satisfy three basic properties.

Property 1. $\bar{H}(x, u, p_1, p_2)$ is consistent with $H(x, u, p)$. That is, $\bar{H}(x, u, p, p) = H(x, u, p)$.

Property 2. $\bar{H}(x, u, p_1, p_2)$ is nonincreasing in the p_1 argument and nondecreasing in the p_2 argument.

Property 3. For all $|p_1| \leqq K$ and $|p_2| \leqq K$, $\bar{H}(x, u, p_1, p_2)$ satisfies Assumption A′ above.

Of course, Property 3 simply says that $\bar{H}(x, u, p_1, p_2)$ is strictly increasing in u at the rate prescribed by Assumption A′. We now give the following theorem.

THEOREM 3. *With Assumptions* A′ *and* B′ *above, suppose that* u^Δ *comes from scheme* (4.3), *where* $\bar{H}(x, u, p_1, p_2)$ *satisfies Properties* 1, 2 *and* 3. *Then,* (4.3) *generates a unique approximate solution* u^Δ, *and moreover* u^Δ *converges to* $u = \lim_\varepsilon u_\varepsilon$ *at least as fast as*

$$|u^\Delta - u|_\infty \leqq \text{const.} \sqrt{\Delta},$$

where above, u *is the viscosity* <u>*limit*</u> *solution of* (4.1) *and* $\Delta = \max_{0 \leqq j \leqq J-1} (x_{j+1} - x_j)$.

Remark 4.1. In fact, the hypotheses of Theorem 3 guarantees that $\lim_\varepsilon u_\varepsilon$ exists. This is seen by checking that in the present situation the derivative of u_ε remains uniformly bounded for $\varepsilon > 0$. The Arzela-Ascoli theorem combined with the results of Proposition 1 should now make the remark obvious.

Before proving Theorem 3, we give two examples of finite difference operators which satisfy Properties 1, 2 and 3. Furthermore, we show that the rate above is the best possible under the hypotheses of Theorem 3.

Example 1. The Lax-Friedrichs difference operator [11], [20], is based upon approximating $H(x, u, du/dx)$ by a convex combination of $H(x, u, D^+u)$ and $H(x, u, D^-u)$ along with the introduction of an artificial numerical viscosity term. To be more specific, $\bar{H}$ is given by

$$\bar{H}(x, u, p_1, p_2) = \theta H(x, u, p_1) + (1 - \theta) H(x, u, p_2) - c(p_1 - p_2),$$

where θ is chosen in $[0, 1]$ and

$$c \geqq \max\,(\theta \sup H_p, (\theta - 1) \inf H_p, 0).$$

Clearly, this difference operator satisfies Properties 1, 2 and 3 above. Moreover, if $H_p \geqq 0$ (resp. $H_p \leqq 0$), we could have chosen $\theta = 0$ (resp. $\theta = 1$), and $c = 0$, thus giving a scheme based on backward (resp. forward) differencing.

Example 2. The Godunov difference operator ([9], [20]) is given by

$$\bar{H}(x, u, p_1, p_2) = \begin{cases} \min\limits_{v \in [p_2, p_1]} H(x, u, v) & \text{if } p_2 < p_1, \\ \max\limits_{v \in [p_1, p_2]} H(x, u, v) & \text{if } p_1 \leqq p_2. \end{cases}$$

This difference operator clearly satisfies Properties 1 and 2, and a straightforward exercise will verify that it satisfies Property 3 as well. Again, when $H(x, u, p)$ is monotone in p, the scheme reduces to either a backward or a forward difference scheme.

Next, we show that the rate of convergence of Theorem 3 is sharp. We again consider the trivial example (3.9) and we approximate its viscosity limit solution ($u=0$) by the Lax–Friedrichs difference scheme; however, we intentionally add too much numerical viscosity (we take $c=1$ rather than the allowable $c=0$). Setting $x_{j+1}-x_j=h$, where $h=1/J$, u_j is required to satisfy:

$$\begin{aligned} &-(D^+u_j-D^-u_j)+u_j=0,\\ &D^-u_0=0, \qquad D^+u_J=1. \end{aligned} \tag{4.4}$$

One easily computes the exact solution of (4.4)

$$u_j=\frac{h}{\alpha_2^{J+1}-\alpha_1^{J+1}}\left[\frac{1}{1-\alpha_1}\alpha_1^{j+1}+\frac{1}{\alpha_2-1}\alpha_2^{j+1}\right],$$

where

$$\alpha_1=1+\frac{h}{2}-\sqrt{h}\cdot\left(1+\frac{h}{4}\right)^{1/2}, \qquad \alpha_2=1+\frac{h}{2}+\sqrt{h}\cdot\left(1+\frac{h}{4}\right)^{1/2},$$

and furthermore, since $u_j\geqq 0$, we have that

$$|u^\Delta-u|_\infty\geqq u_J\geqq\frac{h}{\alpha_2-1}.$$

Finally, calculating the right-hand side above, we arrive at

$$|u^\Delta-u|_\infty\geqq\sqrt{h}\left(\frac{\sqrt{h}}{2}+\left(1+\frac{h}{4}\right)^{1/2}\right)^{-1}=\sqrt{h}+O(h),$$

which is exactly the rate of Theorem 3.

We shall prove Theorem 3 via three lemmas.

LEMMA 2. *Assume that $\bar H(x,u,p_1,p_2)$ satisfies Properties* 1, 2 *and* 3 *above. Then, if the difference scheme* (4.3) *had a solution, say u^Δ, u^Δ is bounded and has a bounded Lipschitz constant, uniformly in $\Delta>0$.*

Proof. We first prove that u^Δ must be uniformly bounded. Suppose that $\max_{0\leqq j\leqq J}u_j\geqq 0$ is attained for some $1\leqq j_0\leqq J-1$. Since at an interior maximum $D^+u_{j_0}\leqq 0\leqq D^-u_{j_0}$, (4.3) and Property 2 imply that

$$\bar H(x_{j_0},u_{j_0},0,0)\leqq\bar H(x_{j_0},u_{j_0},D^+u_{j_0},D^-u_{j_0})=0.$$

Therefore, we have from Property 3 that

$$\mu_0 u_{j_0}\leqq(\max(u_{j_0},1))^{\eta_1}|\bar H(x_{j_0},0,0,0)|.$$

Similarly, if $\max_{0\leqq j\leqq J}u_j\geqq 0$ is attained at $j=0$ or $j=J$, we would have that

$$\mu_{|\gamma_0|}u_0\leqq(\max(u_0,1))^{\eta_1}|\bar H(0,0,0,-\gamma_0)|$$

or

$$\mu_{|\gamma_1|}u_J\leqq(\max(u_J,1))^{\eta_1}|\bar H(1,0,\gamma_1,0)|,$$

which proves that $u^\Delta(x)$ must be bounded above independent of $\Delta>0$. An identical argument would show that $\min_{0\leqq j\leqq J}u_j$ must be bounded below independent of $\Delta>0$.

Next, we show that $|D^+u_j|$ must be uniformly bounded. Suppose that $\max_{0\leqq j\leqq J-1} D^+u_j \geqq \max(-\gamma_0, \gamma_1, 0)$ is attained at j_0. Again using (4.3), we must have that

$$\begin{aligned} 0 &= \bar{H}(x_{j_0+1}, u_{j_0+1}, D^+u_{j_0+1}, D^+u_{j_0}) - \bar{H}(x_{j_0}, u_{j_0}, D^+u_{j_0}, D^+u_{j_0-1}) \\ &\geqq \bar{H}(x_{j_0+1}, u_{j_0+1}, D^+u_{j_0}, D^+u_{j_0}) - \bar{H}(x_{j_0}, u_{j_0}, D^+u_{j_0}, D^+u_{j_0}) \\ &= H(x_{j_0+1}, u_{j_0+1}, D^+u_{j_0}) - H(x_{j_0}, u_{j_0}, D^+u_{j_0}). \end{aligned}$$

Setting $K = D^+u_{j_0} \geqq 0$, we have from above and Property 3 that

$$\mu_K K \leqq \left|\frac{\partial}{\partial x} H(\xi, u_{j_0}, K)\right| (\max(|u_{j_0}|, 1))^{\eta_1},$$

and this inequality combined with Assumption B′ implies that

$$K \leqq (\max(K, 1))^{\eta_2} C(|u_{j_0}|)(\max(|u_{j_0}|, 1))^{\eta_1}.$$

Therefore, D^+u_j is bounded above, again independent of $\Delta > 0$. A similar argument would show that D^+u_j is bounded below independent of $\Delta > 0$. This proves the lemma.

LEMMA 3. *Assume that $\bar{H}(x, u, p_1, p_2)$ satisfies Properties* 1, 2 *and* 3. *Then, the difference scheme* (4.3) *has a unique solution.*

Proof. Consider the map $F_\nu : \mathbf{R}^{J+1} \to \mathbf{R}^{J+1}$, defined by

$$(F_\nu(u))_j = u_j - \nu \bar{H}(x_j, u_j, D^+u_j, D^-u_j), \tag{4.5}$$

for $0 \leqq j \leqq J$, where $-D^-u_0 = \gamma_0$ and $D^+u_J = \gamma_1$. We show below that F_ν has a unique fixed point and obviously this fixed point is the desired solution of difference scheme (4.3). We may assume that $\bar{H}(x, u, p_1, p_2)$ above is globally Lipschitz continuous, since $\bar{H}$ could be modified in a smooth way outside the bounded a priori domain established by the previous lemma.

We now claim that $(F_\nu(u))_j$ is a nondecreasing function in u_{j-1}, u_j and u_{j+1}, provided that ν is chosen sufficiently small. Assume for simplicity that $\bar{H}$ is smooth. We then find upon differentiating that

$$\frac{\partial}{\partial u_{j-1}}(F_\nu(u))_j = \nu \bar{H}_{p_2}/(x_j - x_{j-1}) \qquad \text{for } 1 \leqq j \leqq J,$$

$$\frac{\partial}{\partial u_{j+1}}(F_\nu(u))_j = -\nu \bar{H}_{p_1}/(x_{j+1} - x_j) \qquad \text{for } 0 \leqq j \leqq J-1,$$

and Property 2 implies that these quantities are nonnegative. Furthermore,

$$\frac{\partial}{\partial u_j}(F_\nu(u))_j = 1 - \nu\{\bar{H}_u - \bar{H}_{p_1}/(x_{j+1} - x_j) + \bar{H}_{p_2}/(x_j - x_{j-1})\}, \tag{4.6}$$

for $1 \leqq j \leqq J-1$, and $\bar{H}_{p_2} = 0$ for $j = 0$ and $\bar{H}_{p_1} = 0$ for $j = J$. Therefore, since $\bar{H}$ is assumed to be globally Lipschitz continuous, we can choose ν small enough so that these derivatives are nonnegative as well.

Next, we show that F_ν has the fixed-point property for ν as above (ν should be thought of as an artificial time parameter and the restriction on ν imposed in (4.6) as a CFL condition). Let $u \in \mathbf{R}^{J+1}$ and $v \in \mathbf{R}^{J+1}$ and define $\tau = v - u$. Now consider

$$F_\nu(v) - F_\nu(u) = F_\nu(u + \tau) - F_\nu(u). \tag{4.7}$$

Setting $\tau_M = \max(\max_{0\leqq j\leqq J} \tau, 0)$, we have by the claim above, that

$$(F_\nu(u + \tau) - F_\nu(u))_j \leqq (F_\nu(u + \tau_M \vec{1}) - F_\nu(u))_j. \tag{4.8}$$

Now, recalling the definition of F_ν in (4.5), we see that the right-hand side of (4.8) is equal to

$$\tau_M - \nu(\bar{H}(x_j, u_j + \tau_M, D^+u_j, D^-u_j) - \bar{H}(x_j, u_j, D^+u_j, D^-u_j)),$$

which by Property 3 is bounded above by

$$\tau_M(1 - \nu\hat{\mu}),$$

where $\hat{\mu}$ is the appropriate positive constant of Property 3 governed by the a priori domain of Lemma 2. Setting $\tau_m = \min(\min_{0\leqq j\leqq J} \tau, 0)$ and repeating the argument above, we find that

$$\tau_m(1 - \nu\hat{\mu}) \leqq (F_\nu(v) - F_\nu(u))_j \leqq \tau_M(1 - \nu\hat{\mu}). \tag{4.9}$$

Therefore, the Banach fixed-point theorem guarantees a unique fixed point of F_ν, for ν sufficiently small, which is the desired result.

Remark 4.2. Inequality (4.9) tells us that implementing an artificial time method, $(u^{n+1} = F_\nu(u^n))$, to obtain a solution of difference scheme (4.3), converges at an l^∞ rate of $e^{-\hat{\mu}t}$. This, of course, is computationally slow in light of the increment restriction imposed by (4.6). We recommend a few iterations of artificial time to pull the initial approximation into the l^∞ domain of attraction for Newton's method, which with some "smoothness," converges at a much faster quadratic rate.

The next lemma is crucial to establish the fact that u^Δ satisfies the approximate viscosity inequalities.

LEMMA 4. *Suppose* $\phi(x) = \kappa|x - y|^2 + \psi(x)$, *where* $y \in [0, 1]$ *is fixed,* κ *is a constant and* $\psi(x)$ *is an affine function with* $-\psi'(0) = \gamma_0$ *and* $\psi'(1) = \gamma_1$. *Then*:

(a) *If* $\kappa \geqq 0$ *and* $\max_{x\in[0,1]}(u^\Delta(x) - \phi(x))$ *is attained for some* $\xi \neq x_j$, $0 \leqq j \leqq J$, *we have* $D^+u_{j_0} - D^-u_{j_0} \leqq \Delta \cdot \phi_{xx}(\xi)$, *where* x_{j_0} *is the nearest grid point to* ξ.

(b) *If* $\kappa \leqq 0$ *and* $\min_{x\in[0,1]}(u^\Delta(x) - \phi(x))$ *is attained for some* $\xi \neq x_j$, $0 \leqq j \leqq J$, *we have* $D^+u_{j_0} - D^-u_{j_0} \geqq \Delta \cdot \phi_{xx}(\xi)$, *where* x_{j_0} *is the nearest grid point to* ξ.

Proof. We prove (a) only since the proof of (b) is identical. Let x_{j_0} be the nearest grid point to ξ. We have three basic cases to examine: $x_{j_0} = 0$, $x_{j_0} = 1$ and $0 < x_{j_0} < 1$.

Case 1. When $x_{j_0} = 0$, we must have $D^+u_0 = \phi_x(\xi)$ since $\xi \in (0, x_1)$ is where the maximum of $u^\Delta - \phi$ occurs. However, because ϕ is quadratic, $\phi_x(\xi) = \phi_x(0) + \xi\phi_{xx}(\xi)$. Therefore, $D^+u_0 = \xi\phi_{xx}(\xi) + \phi_x(0) \leqq \xi\phi_{xx}(\xi) - \gamma_0$.

Case 2. The case when $x_{j_0} = 1$ is identical to Case 1 above.

Case 3. Suppose now that $\xi \in (x_{j_0-1}, x_{j_0})$ and choose an arbitrary $\tau \in (x_{j_0}, x_{j_0+1})$ (if, on the other hand, $\xi \in (x_{j_0}, x_{j_0+1})$ the argument below is essentially the same). Using the definition of u^Δ and the fact that $(u^\Delta - \phi)(\xi)$ is maximum, we have that

$$\begin{aligned} &u_{j_0} + D^-u_{j_0}(\xi - x_{j_0}) - \phi(\xi) \geqq u_{j_0} + D^+u_{j_0}(\tau - x_{j_0}) - \phi(\tau), \\ &D^-u_{j_0} = \phi_x(\xi). \end{aligned} \tag{4.10}$$

Therefore, a simple calculation will show that (4.10) implies

$$D^+u_{j_0} - D^-u_{j_0} \leqq \frac{\phi_x(\xi)(\xi - \tau) + \phi(\tau) - \phi(\xi)}{\tau - x_{j_0}}. \tag{4.11}$$

Taylor's theorem allows us to write the right-hand side of (4.11) as

$$\frac{1}{2}\left[\frac{(\tau - \xi)^2}{\tau - x_{j_0}}\right]\phi_{xx}(\xi).$$

Recall that we have assumed our grid satisfies the constraint $(x_{j+1}-x_j)\geqq\frac{1}{2}(x_j-x_{j-1})$. This allows us to minimize the bracketed term above by choosing $\tau=2x_j-\xi$. Doing this, we have

$$D^+u_{j_0}-D^-u_{j_0}\leqq 2(x_{j_0}-\xi)\phi_{xx}(\xi),$$

and since x_{j_0} is the *nearest* grid point to ξ, the proof is complete.

Proof of Theorem 3. The proof of the theorem is complete (Lemmas 2 and 3), except for showing that u^Δ satisfies the approximate viscosity inequalities of Definition 2. With this in mind, set $\phi_1(x)=|x-y_0|^2/\delta+\psi(x)$, where $y_0\in[0,1]$, is fixed and $\psi(x)$ is affine, with $\psi'(0)=-\gamma_0$ and $\psi'(1)=\gamma_1$ (as in Definition 2(a)). Suppose now that $u^\Delta-\phi_1$ is maximum at $\xi\in[0,1]$. To show that u^Δ is an approximate viscosity subsolution, we must verify that

$$H(\xi, u^\Delta(\xi), \phi_{1x}(\xi))\leqq(K\cdot\Delta)\cdot\phi_{1xx}(\xi)+C\Delta, \tag{4.12}$$

where K is some constant, independent of Δ, and as always in this section, $\Delta=\max_{0\leqq j\leqq J-1}(x_{j+1}-x_j)$.

Using difference scheme (4.3) and Property 1, we have that for every $0\leqq j\leqq J$

$$H(\xi, u^\Delta(\xi), \phi_{1x}(\xi))=\bar H(\xi, u^\Delta(\xi), \phi_{1x}(\xi), \phi_{1x}(\xi))-\bar H(x_j, u_j, D^+u_j, D^-u_j),$$

and we rewrite this identity as

$$\begin{aligned}H(\xi, u^\Delta(\xi), \phi_{1x}(\xi))=&[\bar H(x_j, u_j, \phi_{1x}(\xi), \phi_{1x}(\xi))-\bar H(x_j, u_j, D^+u_j, D^-u_j)]\\&+[\bar H(\xi, u^\Delta(\xi), \phi_{1x}(\xi), \phi_{1x}(\xi))\\&\quad-\bar H(x_j, u_j, \phi_{1x}(\xi), \phi_{1x}(\xi))].\end{aligned} \tag{4.13}$$

The second term on the right-hand side of (4.13) is bounded above by

$$L_x|\xi-x_j|+L_uL|\xi-x_j|, \tag{4.14}$$

where L_x and L_u are the Lipschitz constants of $\bar H$ in the x and u arguments, respectively, and L is the Lipschitz constant of u^Δ. The first term on the right-hand side of (4.13) can be written as

$$\bar H_{p_1}\cdot(\phi_{1x}(\xi)-D^+u_j)+\bar H_{p_2}\cdot(\phi_{1x}(\xi)-D^-u_j), \tag{4.15}$$

where, again, we have assumed that $\bar H$ is smooth for simplicity.

If $\xi=x_{j_0}$ for some $0\leqq j_0\leqq J$, we have nothing to prove since it is an easy exercise to determine that in this case $D^+u_{j_0}\leqq\phi_{1x}(\xi)\leqq D^-u_{j_0}$ when x_{j_0}, $(=\xi)$, is a maximizer of $u^\Delta-\phi_1$. (Recall by the definition of ϕ_1 that $\phi_1(0)\leqq D^-u_0$ and $\phi_{1x}(1)\geqq D^+u_J$ in the event that $\xi=0$ or 1.) Therefore setting $j=j_0$ in (4.15) and recalling Property 2 (which says that $\bar H_{p_1}\leqq 0\leqq\bar H_{p_2}$), verifies the approximate viscosity inequality of Definition 2(a) here in a trivial way.

If, on the other hand, $\xi\neq x_j$ for all $0\leqq j\leqq J$, take x_{j_0} to be the nearest grid point to ξ. Set $j=j_0$ in (4.15) and insert the identity $\phi_{1x}(\xi)=D^-u_{j_0}$, (or $\phi_{1x}(\xi)=D^+u_{j_0}$) into it. Using the result of Lemma 4 allows us to combine (4.15) with (4.13) to arrive at

$$H(\xi, u^\Delta(\xi), \phi_{1x}(\xi))\leqq(K\cdot\Delta)\cdot\phi_{1xx}(\xi)+C\Delta,$$

where $K=\max(-H_{p_1}, H_{p_2})$ and C is given by $\frac{1}{2}(L_x+L_uL)$.

An identical argument will show that u^Δ is an approximate viscosity supersolution, (see Definition 2(b)), and so by applying the abstract result of Theorem 2, the proof of Theorem 3 is complete.

REFERENCES

[1] R. F. ANDERSON AND S. OREY, *Small random perturbations of dynamical systems with reflecting boundary*, Nagoya Math. J., 60 (1976), pp. 189-216.

[2] C. BARDOS, A. Y. LEROUX AND J. C. NEDELEC, *First order quasilinear equations with boundary conditions*, Comm. Partial Differential Equation, 4 (1979), pp. 1017-1034.

[3] I. CAPPUZZO DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, to appear.

[4] M. G. CRANDALL AND P. L. LIONS, *Conditions d'unicité pour les solutions généralisées des équations de Hamilton-Jacobi du premier ordre*, C. R. Acad. Sci. Paris, 292 (1981), pp. 183-186.

[5] ———, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[6] ———, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1-19.

[7] M. G. CRANDALL AND P. E. SOUGANIDIS, *Developments in the theory of nonlinear first-order partial differential equations*, in Differential Equations, I. W. Knowles and R. T. Lewis, eds., Elsevier Science, North-Holland, 1984, pp. 131-142.

[8] M. G. CRANDALL AND R. NEWCOMB, *Viscosity solutions of Hamilton-Jacobi equations of the boundary*, to appear.

[9] S. K. GODUNOV, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 271-290. (In Russian.)

[10] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1977.

[11] P. D. LAX, *Shock waves and entropy*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 603-634.

[12] P. L. LIONS, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Parts* 1 *and* 2, Comm. Partial Differential Equations, 8 (1983), pp. 1101-1174, pp. 1229-1276.

[13] ———, *Résolution de problémes elliptiques quasilineaires*, Arch. Rational Mech. Anal., 74 (1980), pp. 335-353.

[14] ———, *Quelques remarques sur les problémes eilliptiques quasilineaires du second ordre*, to appear.

[15] ———, *Neumann type boundary conditions for Hamilton-Jacobi equations*, to appear.

[16] P. L. LIONS AND B. PERTHAME, *Quasi-variational inequalities and ergodic impulse control*, to appear.

[17] P. L. LIONS AND A. S. SZNITMAN, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 34 (1984), pp. 511-537.

[18] J. LORENZ AND R. SANDERS, *Second order nonlinear singular perturbation problems with boundary conditions of mixed type*, this Journal, 17 (1986), pp. 580-594.

[19] ———, *On the rate of convergence of viscosity solutions for boundary value problems*, this Journal, 18 (1987), pp. 306-320.

[20] R. SANDERS, *On monotone finite difference schemes with variable spatial differencing*, Math. Comp., 40 (1983), pp. 91-106.

[21] P. E. SOUGANIDIS, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, MRC TSR 2511, University of Wisconsin, Madison, WI, 1982.

[22] ———, *Existence of viscosity solutions of Hamilton-Jacobi equations*, to appear.

[23] J. SERRIN, *The problem of Dirichlet for quasilinear elliptic differential equations with many independent variables*, Philos. Trans. Roy. Soc. London, Ser. A., 204 (1969), pp. 413-469.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

STABILITY AND INSTABILITY FOR SOLUTIONS OF BURGERS' EQUATION WITH A SEMILINEAR BOUNDARY CONDITION*

HOWARD A. LEVINE†

Abstract. In this paper, we present several results concerning the long-time behavior of positive solutions of Burgers' equation $u_t = u_{xx} + \varepsilon u u_x$, $0 < x < 1$, $t > 0$, $u(x, 0)$ given, subject to one of two pairs of boundary conditions: (A) $u(0, t) = 0$, $u_x(1, t) = au^p(1, t)$, $t > 0$, or (B) $u(1, t) = 0$, $u_x(0, t) = -au^p(0, t)$, where $0 < p < \infty$. A complete stability-instability analysis is given. It is shown that some solutions can blow up in finite time. Generalizations replacing $\varepsilon u u_x$ by $(f(u))_x$ and au^p by $g(u)$ are discussed.

Key words. Burgers' equation, stability, instability

AMS(MOS) subject classifications. 35K05, 35K20, 35K55, 35K60, 76E99

1. Introduction. In this paper, we consider two nonstandard initial-boundary value problems for Burgers' equation, namely

$$\text{(A)}\qquad \begin{aligned} &u_t = u_{xx} + \varepsilon u u_x && \text{on } (0, 1) \times (0, \infty),\\ &u_x(1, t) = au^p(1, t) && \text{on } (0, \infty),\\ &u(0, t) = 0 && \text{on } (0, \infty),\\ &u(x, 0) = u_0(x) \text{ prescribed} && \text{on } [0, 1] \end{aligned}$$

and

$$\text{(B)}\qquad \begin{aligned} &u_t = u_{xx} + \varepsilon u u_x && \text{on } (0, 1) \times (0, \infty),\\ &u(1, t) = 0 && \text{on } (0, \infty),\\ &-u_x(0, t) = au^p(0, t) && \text{on } (0, \infty),\\ &u(x, 0) = u_0(x) && \text{on } [0,1]. \end{aligned}$$

Here $p > 0$, ε, $a > 0$, while u^p is defined as $|u|^{p-1}u$. We observe that in this case $\tilde{u}(x, t) = -u(1 - x, t)$ defines a one-to-one, onto correspondence between the solutions of (A) and those of (B). This observation permits us to construct all the stationary solutions of (A) (or (B)) for all real ε, if we know only the positive stationary solutions of (A) and (B) for $\varepsilon \geqq 0$. (Nontrivial stationary solutions of (A) and (B) are necessarily of one sign.)

Our interest in these problems is twofold. First, when $\varepsilon = 0$, (A) and (B) are essentially the same problem. They have been studied from the point of view of potential well-theory (in several space dimensions) in a recent series of papers [6], [7]. The arguments used therein establish the existence of a potential well for which solutions starting in the well remain in the well and for which solutions starting in the exterior of the well are unstable and, indeed, fail to exist for all time. However, when $\varepsilon \neq 0$, such arguments, which demand the existence of a potential energy functional, cannot be applied to problems (A) and (B), for which no such functional exists.

* Received by the editors August 25, 1986; accepted for publication (in revised form) April 6, 1987. This research was sponsored by the U.S. Air Force Office of Scientific Research, Air Force Systems Command, under grant 84-0252.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

Second, in [1], [15], the authors have obtained partial results and done some numerical experiments for

$$\text{(C)}\qquad \begin{aligned} &u_t = u_{xx} + \varepsilon u u_x + a u^p && \text{in } (0,1)\times(0,\infty),\\ &u(0,t) = u(1,t) = 0 && \text{on } (0,\infty),\\ &u(x,0) \text{ prescribed} && \text{on } [0,1]. \end{aligned}$$

It has been observed in [6], [7] that with $\varepsilon = 0$, potential well theory for (C) closely parallels that for (A) (even in several space dimensions). We might therefore expect that when $\varepsilon > 0$, the study of (A) or (B) might provide additional insight into the behavior of solutions of (C).

Although this is true in some generalized sense, the analysis of the bifurcation diagram for (C) is much less well understood than those for (A) or (B). However, numerical calculations show that it is closer to (B) than to (A) in structure.

Our results are in the spirit of the framework considered by Hirsch [3] and Matano [9], [10] for strongly order preserving systems. However, application of their general results to our problem is complicated by the presence of the nonlinear term in the boundary condition. Also we make very strong use of the qualitative dependence of the stationary solutions upon ε, which is probably special to the one space dimensional character of our problem. We hope to pursue this matter in a later work.

Some (but not all) of our local existence results have been obtained by Amann [14], in a more general setting. However, we include these proofs here to make our work self-contained.

The plan of the paper is as follows. In § 2, we characterize the set of nonnegative stationary solutions for a generalization of (A), (B). We then obtain the set of stationary solutions of (A), (B) and give the bifurcation diagrams. Verification of the nature of the diagrams is given in Appendix I. In the third section we examine the questions of stability and unstability of the set of stationary solutions. Finally we briefly discuss the question of local existence and continuation in § 4.

2. Stationary solutions. Here we consider stationary solutions for

$$(\text{A}_1)\qquad \begin{aligned} &u_t = u_{xx} + (f(u))_x && \text{on } (0,1)\times(0,\infty),\\ &u_x(1,t) = g(u(1,t)) && \text{on } (0,\infty),\\ &u(0,t) = 0 && \text{on } (0,\infty),\\ &u(x,0) = u_0(x) && \text{on } [0,1] \end{aligned}$$

and

$$(\text{B}_1)\qquad \begin{aligned} &u_t = u_{xx} + (f(u))_x && \text{on } (0,1)\times(0,\infty),\\ &u(1,t) = 0 && \text{on } (0,\infty),\\ &-u_x(0,t) = g(u(0,t)) && \text{on } (0,\infty),\\ &u(x,0) = u_0(x) && \text{on } [0,1], \end{aligned}$$

where f, g are real valued, continuously differentiable functions defined on R^1 with $f(0) = g(0) = 0$ and where $ug(u) > 0$ if $u \neq 0$. We will impose additional hypotheses below. However, these will include the choice $f(u) = \varepsilon u^2/2$, $g(u) = a|u|^{p-1}u$ with $p > 0$. We shall focus on the behavior of nonnegative solutions of (A_1), (B_1) and their corresponding stationary problems.

The following lemma is a simple consequence of the first and second maximum principles for elliptic equations. No particular sign assumptions need be placed on f', f'', g.

LEMMA 2.1. *Let f be twice continuously differentiable. Nonzero stationary solutions of* (A_1) *and* (B_1) *cannot change sign. Positive solutions $w(x)$ of* (A_1) *satisfy* $w'(x)>0$ *on* $[0,1]$, *while positive solutions of* (B_1) *satisfy* $w'(x)<0$ *on* $[0,1]$.

Proof. The first statement follows from the maximum principle. For the first part of the second statement, we must have $w'(0)\geqq 0$. This inequality is strict unless $w\equiv 0$. If w' changed sign on $[0,1)$, w would have an interior maximum which cannot happen unless $w=$ constant $=0$. If $w'(1)=0$, then, from the Hopf second principle, $w(x)\equiv w(1)$ and consequently $w(x)\equiv 0$. If, for some $x_0\in(0,1)$ $w'(x_0)=0$ and $w'(x)\geqq 0$ otherwise, then $w''(x_0)=0$ also. However $v=w'$ then satisfies $v''+f'(w)v'+f''(w)v^2=0$ with $v(x_0)=v'(x_0)=0$. By uniqueness, $v\equiv 0$ and again $w(x)\equiv w(1)$.

A similar argument holds for the second part of the second statement. □

THEOREM 2.1A. *Let $f'>0$ for $u>0$. Let $w(x)$ be a positive stationary solution of* (A_1), C^2 *on* $(0,1)$ *and* C^1 *on* $[0,1]$. *Let $w_1\equiv w(1)$. Then*

$$\int_0^{w(x)}\frac{d\sigma}{g(w_1)+f(w_1)-f(\sigma)}=x \tag{2.1}$$

for $0\leqq x\leqq 1$. Conversely, if $w_1>0$ solves

$$\int_0^{w_1}\frac{d\sigma}{g(w_1)+f(w_1)-f(\sigma)}=1, \tag{2.2}$$

and w solves (2.1) *with this degree of smoothness, with $w(1)=w_1$, then w is a positive stationary solution of* (A_1).

THEOREM 2.1B. *Let $f\geqq 0$ for $u\geqq 0$. Let $w(x)$ be a positive stationary solution of* (B_1), C^2 *on* $(0,1)$ *and* C^1 *on* $[0,1]$. *Let $w_0=w(0)$. Then*

$$\int_0^{w(x)}\frac{d\sigma}{g(w_0)-f(w_0)+f(\sigma)}=1-x \tag{2.3}$$

for $0\leqq x\leqq 1$ and $g(w_0)-f(w_0)>0$. Conversely, if $w(0)=w_0>0$, $g(w_0)-f(w_0)>0$, w_0 solves

$$\int_0^{w_0}\frac{d\sigma}{g(w_0)-f(w_0)+f(\sigma)}=1, \tag{2.4}$$

w solves (2.3) *with this degree of smoothness, and $w(0)=w_0$, then w is a positive stationary solution of* (B_1).

Proof. To prove the first of these, we note that (2.1) and (2.2) follow from

$$w''(x)+(f(w(x)))'=0, \qquad 0<x<1,$$

$$w(0)=-w'(1)+g(w(1))=0,$$

after noting that $w'(x)=-f(w(x))+f(w_1)+g(w_1)>0$ and a second quadrature.

For the converse, we observe that if $w(\cdot)$ satisfies (2.1), then $w(x)<w_1$ if $x<1$. To see this, suppose that $w(\bar{x})\geqq w_1$ for some $\bar{x}\in[0,1)$. If $h(\sigma)\equiv g(w_1)+f(w_1)-f(\sigma)$ has no roots, then $h(\sigma)>0$ and

$$1=\int_0^{w_1}\frac{d\sigma}{h(\sigma)}\leqq\int_0^{w(\bar{x})}\frac{d\sigma}{h(\sigma)}=\bar{x}$$

so that $\bar{x}=1$. If $h(\sigma)$ has a root, say $\bar{\sigma}$, then this root is unique. Moreover, $\bar{\sigma}>\sup\{w(x)\,|\,0\leqq x\leqq 1\}$; otherwise there would be $\bar{x}\in(0,1)$ with $\bar{\sigma}=w(\bar{x})$. But $h(\sigma)=f(\bar{\sigma})-f(\sigma)\approx f'(\bar{\sigma})(\bar{\sigma}-\sigma)$ so that

$$\bar{x}=\int_0^{w(\bar{x})}\frac{d\sigma}{h(\sigma)}=\int_0^{\bar{\sigma}}\frac{d\sigma}{f(\bar{\sigma})-f(\sigma)}=+\infty.$$

Thus $w(x)\leqq\bar{\sigma}-\delta$ for some $\delta>0$ and all $x\in[0,1]$. Therefore we may differentiate (2.1) to find that $w'(x)+f(w(x))=g(w_1)+f(w_1)$ and $w''(x)+(f(w(x)))'=0$. Thus $w'(1)=g(w_1)$. Also, if $\bar{x}\in[0,1)$ is such that $w(1)<w(\bar{x})<\bar{\sigma}$, then by the argument above we find again that $\bar{x}\geqq 1$, which is impossible. Therefore $w(x)\leqq w_1$ and thus $w'(x)\geqq f(w_1)-f(w(x))\geqq 0$. Thus $w(0)\equiv\lim_{x\to 0^+}w(x)$ exists and, by (2.1), must be zero. □

The proof of Theorem 2.1B is similar. We note that if w is a stationary solution, then the conservation law $w'(x)+f(w(x))=$ constant yields (since $f(0)=0$), $f(w(0))-g(w(0))=w'(1)$ which is negative by the lemma.

Equations (2.2) and (2.4) can have several solutions, each corresponding to a stationary solution. With somewhat further restrictions on f, g we can prove that these solutions are ordered.

THEOREM 2.2A. *Let u_1, v_1 be solutions of* (2.2) *with $u_1>v_1>0$. Let $u(x)$, $v(x)$ be the corresponding solutions of* (2.1) *with $u(1)=u_1$, $v(1)=v_1$. Suppose that $f(u_1)+g(u_1)>f(v_1)+g(v_1)$ (which holds if $f+g$ is strictly increasing). Then $u(x)>v(x)$ for x in* $(0,1]$.

THEOREM 2.2B. *Let u_0, v_0 be solutions of* (2.4) *with $u_0>v_0>0$, $g(u_0)-f(u_0)>0$, $g(v_0)-f(v_0)>0$ and let $u(x)$, $v(x)$ denote the corresponding solutions of* (2.3). *If f' is strictly increasing, then $u(x)>v(x)$ on* $[0,1)$.

Theorem 2.2B is an easy consequence of the maximum principle. If $w(x)=u(x)-v(x)$, then $w(0)>0$, $w(1)=0$ and, on $(0,1)$,

$$w''+f'(u)w'+(f'(u)-f'(v))v'=0. \tag{2.5}$$

Since $v'<0$ and f' is strictly increasing, the usual arguments show that w cannot have an interior negative minimum. Therefore $w\geqq 0$. If w had an interior zero at x_0, it would also have a positive maximum on $(x_0,1)$. However from (2.5) we see that this is false also.

(The hypotheses on f, g *do not* imply that there are any solutions at all of (2.2), (2.4).)

To prove Theorem 2.2A, we see from the conservation laws that for any $x\in(0,1)$

$$u_x(0)=g(u_1)+f(u_1)=u_x(x)+f(u(x))$$

and

$$v_x(0)=g(v_1)+f(v_1)=v_x(x)+f(v(x)).$$

From the hypothesis we find $u_x(0)>v_x(0)$. Since $u(0)=v(0)=0$, $u(x)>v(x)$ in a neighborhood of $x=0$. If $\bar{x}$ is the first point in $(0,1]$ where $u(\bar{x})=v(\bar{x})$, we see from the above that $u_x(\bar{x})>v_x(\bar{x})$. This inequality holds in a left open neighborhood of $\bar{x}$, say $(\bar{x}-\delta,\bar{x}]$. But then we obtain a contradiction from

$$0=u(\bar{x})-v(\bar{x})=\int_{\bar{x}-\delta}^{\bar{x}}(u_x(x)-v_x(x))\,dx+[u(\bar{x}-\delta)-v(\bar{x}-\delta)]. \tag{2.6}$$

□

Example 2.1. $f(u)=\frac{1}{2}\varepsilon u^2$, $g(u)=a|u|^{p-1}u$, a, $\varepsilon>0$. In this case, (2.2) is equivalent to

$$F(w_1)\equiv\int_0^1\frac{d\sigma}{(2a/\varepsilon)w_1^{p-2}+1-\sigma^2}=\frac{1}{2}\varepsilon w_1. \tag{2.7}$$

We find that for $p=2$ there is one solution for all $\varepsilon>0$. For $p>2$, $F'(w_1)<0$, $F(w_1)\to+\infty$ as $w_1\downarrow 0$ and $F(w_1)\to 0$ as $w_1\to+\infty$ so that there is only one solution for all $\varepsilon>0$ in this case also. If $1<p<2$, we set $v_1=(2a/\varepsilon)w_1^{p-2}$, $\delta=a^{1/(2-p)}(\varepsilon/2)^{(p-1)/(p-2)}$ and seek the number of positive solutions of

$$Q(v_1)\equiv v_1\int_0^1\frac{d\sigma}{v_1+(1-\sigma^2)}=\delta v_1^{(p-1)/(p-2)}\equiv R(v_1).$$

It is easy to see that $Q(0)=0$, $Q'(v_1)>0$, $Q(+\infty)=1$ while $R'(v_1)<0$, $R(v_1)\to 0$ as $v_1\to+\infty$, $R(v_1)\to+\infty$ as $v_1\to 0^+$. Therefore problem (A) has exactly one positive stationary solution for all $a>0$, $\varepsilon>0$ in this case also.

The case $p=1$ must be treated separately. We observe that when $p=1$, we must solve $Q(v)=a$. Therefore there is one and only one positive stationary solution when $0<a<1$ and none when $a\geqq 1$.

The case $0<p<1$ is more difficult. We try to solve

$$\phi(v_1)=v_1^{1/(2-p)}\int_0^1\frac{d\sigma}{v_1+(1-\sigma^2)}=\delta$$

where, after an abuse of notation, we let

$$\phi(\alpha)=\tfrac{1}{2}(\alpha^2-1)^q\alpha^{-1}\ln((\alpha+1)/(\alpha-1))$$

on $(1,\infty)$, with $q=1/(2-p)$ and $\alpha=(v_1+1)^{1/2}$. It is easily seen that $\frac{1}{2}<q<1$ and that

$$\phi'(\alpha)=\tfrac{1}{2}(\alpha^2-1)^{q-1}\alpha^{-2}K(\alpha)$$

where

$$K(\alpha)=[(2q-1)\alpha^2+1]\ln((\alpha+1)/(\alpha-1))-2\alpha$$

and that

$$\tfrac{1}{2}K'(\alpha)=(2q-1)\alpha\ln((\alpha+1)/(\alpha-1))-2q\alpha^2/(\alpha^2-1).$$

We see that, as $\alpha\to+\infty$, $K'(\alpha)/2\alpha\sim 2(q-1)\alpha/(\alpha^2-1)<0$ and that

$$\left(\frac{K'(\alpha)}{2\alpha}\right)'=\frac{1}{(\alpha^2-1)^2}[2(1-q)\alpha^2+(6q-2)]$$

which is positive. Therefore K' is negative on $(1,\infty)$. However $\lim_{\alpha\to 1^+}K(\alpha)=+\infty$ while $K(\alpha)\sim 4(q-1)\alpha$ (<0) as $\alpha\to\infty$. Therefore K has exactly one sign change and ϕ first increases and then decreases on $(1,\infty)$. We note also that $\lim_{\alpha\to 1^+}\phi(\alpha)=0$ and, by L'Hopital's rule $\phi(\alpha)\approx(2/(2q-1))\alpha^{-2(1-q)}$ as $\alpha\to+\infty$ so that $\lim_{\alpha\to+\infty}\phi(\alpha)=0$. Thus the equation $\phi(\alpha)=\delta$ has zero, one or two solutions accordingly as

$$\varepsilon>2(\bar{\phi})^{(p-2)/(p-1)}a^{1/(p-1)},\qquad \varepsilon=2(\bar{\phi})^{(p-2)/(p-1)}a^{1/(p-1)}$$

or

$$\varepsilon<2(\bar{\phi})^{(p-2)/(p-1)}a^{1/(p-1)}$$

where

$$\bar{\phi}=\max_{1<\alpha<\infty}\phi(\alpha).$$

For (2.4) we have, in this case,

$$F(w_0)=\int_0^1\frac{d\sigma}{(2a/\varepsilon)w_0^{p-2}-1+\sigma^2}=\frac{1}{2}\varepsilon w_0 \tag{2.8}$$

with the additional condition that

$$(2a/\varepsilon)w_0^{p-2} > 1.$$

In this case, it is convenient to define

$$\beta = (2a/\varepsilon)w_0^{p-2} - 1$$

and seek positive solutions of

$$G(\beta) \equiv \int_0^1 \frac{d\sigma}{\beta + \sigma^2} = \delta(\beta+1)^{1/(p-2)} \equiv H(\beta) \tag{2.9}$$

with δ as above. If $p \geqq 2$, we see as above that there is exactly one positive solution of (2.9) and hence (2.8). On the other hand, if $1 < p < 2$, we examine the equation (with $\alpha^2 = \beta$, $\alpha > 0$)

$$I(\alpha) \equiv \frac{(\alpha^2+1)^{1/(2-p)}}{\alpha} \tan^{-1}\left(\frac{1}{\alpha}\right) = \delta.$$

We see that with $q = 1/(2-p)$,

$$I'(\alpha) = (\alpha^2+1)^{q-1}\alpha^{-2}K(\alpha)$$

where

$$K(\alpha) = ((2q-1)\alpha^2 - 1)\tan^{-1}(1/\alpha) - \alpha.$$

We have

$$K'(\alpha) = 2(2q-1)\alpha \tan^{-1}(1/\alpha) - 2q\alpha^2/(\alpha^2+1).$$

Moreover, $K'(\alpha) > 0$ on $(0, \infty)$ while

$$K(\alpha) \to \frac{-\pi}{2} \quad \text{as } \alpha \to 0^+$$

and

$$K(\alpha) \sim [2(p-1)/(2-p)]\alpha$$

as $\alpha \to +\infty$. Therefore $I'(\alpha)$ changes sign exactly once on $(0, \infty)$, $I(\alpha) \to +\infty$ as $\alpha \to 0^+$ and as $\alpha \to +\infty$. Therefore (2.9) has zero, one or two solutions according to whether

$$\varepsilon < 2(\bar{I})^{-(2-p)/(p-1)}a^{1/(p-1)}, \qquad \varepsilon = 2(\bar{I})^{-(2-p)/(p-1)}a^{1/(p-1)}$$

or

$$\varepsilon > 2(\bar{I})^{-(2-p)/(p-1)}a^{1/(p-1)}$$

where

$$\bar{I} = \min_{0<\alpha<\infty} I(\alpha).$$

When $p = 1$, the situation is somewhat different than the case $p > 1$. We must solve (with $\delta = a$ when $p = 1$)

$$J(\alpha) \equiv \tan^{-1}(1/\alpha) - \delta\alpha/(\alpha^2+1) = 0.$$

It is easy to see that when $\delta \in (0, 1]$, $J'(\alpha) < 0$ and the range of J is $(0, \pi/2)$ so that we have no positive stationary solutions in this case. If $\delta > 1$, then $J'(\alpha)$ has a unique positive root at $\bar{\alpha} = [(\delta+1)/(\delta-1)]^{1/2}$ while $J'(\alpha) > 0$ if $\alpha > \bar{\alpha}$ and $J'(\alpha) < 0$ if $\alpha < \bar{\alpha}$.

Since $\bar{\alpha}$ corresponds to a negative minimum of J, we see that there is a unique solution of $J(\alpha)=0$ in $(0, \bar{\alpha})$ and none on $[\bar{\alpha}, \infty)$. Thus, when $a \geqq 1$ there is a unique positive stationary solution for all $\varepsilon > 0$. Otherwise there is none.

In the case $0<p<1$, $K(\alpha) \to -\infty$ as $\alpha \to +\infty$. We write $K'(\alpha) = \alpha Q(\alpha)$ where

$$Q'(\alpha) = \frac{2}{(\alpha^2+1)^2}[(1-q)\alpha^2 - (3q-1)].$$

We see that Q' changes sign from $-(6q-2)$ near $\alpha = 0$ to nearly $2(1-q)\alpha^2/(\alpha^2+1)^2$ for α large. Therefore since $Q(\alpha) \to (2q-1)\pi$ as $\alpha \to 0^+$ and zero at $\alpha = +\infty$, Q changes sign exactly once and hence so does $K'(\alpha)$. The unique root of K' will correspond to a maximum of $K(\alpha)$. Calling this root $\bar{\alpha}$, we have

$$(2q-1)\tan^{-1}(1/\bar{\alpha}) = q\bar{\alpha}/(\bar{\alpha}^2+1).$$

We find

$$K(\bar{\alpha}) = \frac{-1}{(2q-1)(\bar{\alpha}^2+1)} \cdot [(1-q)(2q-1)\bar{\alpha}^3 + (3q-1)\bar{\alpha}]$$

which is negative.

Therefore $K(\alpha)<0$ and $I'(\alpha)<0$. Since $I(\alpha) \sim \alpha^{-2(1-q)}$ as $\alpha \to +\infty$, we see that in this case $I(\alpha) = \delta$ has exactly one solution when $0<p<1$.

The bifurcation diagrams then have the form indicated in Figs. 1.1, 1.2, 2.1 and 2.2. In Appendix II we establish the qualitative shapes of the curves in Figs. 1.1, 1.2, 2.1 and 2.2.

3. Stability-instability-global nonexistence. Here we examine the questions of stability and instability for the time dependent problems (A_1), (B_1) with particular attention focused on (A), (B). We shall assume all solutions are C^2 in x and C^1 in t on $(0, 1)\times(0, T) \equiv D_T$ and continuous in the parabolic cylinder $[0, 1]\times[0, T) \equiv D_T \cup \Gamma_T$. We shall assume f is C^2, for convenience. (See Appendix I.)

LEMMA 3.1A. *Suppose that f' is increasing and that either $g(u)/u$ is bounded in a neighborhood of $u=0$ or else that $u>0$ on $\{1\}\times[0, T)$. Let $u(x, t)$ solve (A_1). If $u(x, 0)>0$ on $(0, 1]$, then $u>0$ on $D_T \cup \Gamma_T$ except at $x=0$. Suppose also that $g(u)/u$ is increasing on $(0, \infty)$. If $w(x)$ is a positive stationary solution of (A_1), $\sigma \in [0, 1)$, and $u(x, 0) \leqq (1-\sigma)w(x)$, then $u(x, t) \leqq (1-\sigma)w(x)$. If $u(x, 0) \geqq (1+\sigma)w(x)$, for some $\sigma>0$, then $u(x, t) \geqq (1+\sigma)w(x)$ on $D_T \cup \Gamma_T$.*

Proof. If u had a negative minimum in $\bar{D}_{T-\delta}$ for some $\delta > 0$, then for any $\lambda, \mu > 0$, $v = e^{-(\lambda x+\mu t)}u$ also would have a negative minimum in $\bar{D}_{T-\delta}$. We choose λ so large that

$$\lambda > \sup\{g(u(1, t))/u(1, t) \,|\, 0 \leqq t \leqq T-\delta\}$$

and then choose μ so large that

$$\mu > \lambda^2 + \lambda \sup\{f'(u(x, t)) \,|\, (x, t) \in \bar{D}_{T-\delta}\}.$$

Then for v we have, in $D_{T-\delta}$,

$$v_t = v_{xx} + (2\lambda + f'(u))v_x + (\lambda f'(u) + \lambda^2 - \mu)v$$

while

$$v_x = (g(u)/u - \lambda)v$$

when $x=1$ and $0<t \leqq T-\delta$. From the first of these, a negative minimum cannot occur in $D_{T-\delta}$ or at $t = T-\delta$ and $0<x<1$, while from the second it cannot occur on $x=1$, $0<t \leqq T-\delta$. Since it cannot occur at $x=0$, we have $u(x, t)>0$ in $\bar{D}_{T-\delta}$ except at $x=0$.

To prove the second statement, we let $v(x)=(1-\sigma)w(x)$ and note that

$$\begin{aligned} v_{xx}+f'(v)v_x &\leqq (1-\sigma)[w_{xx}+f'((1-\sigma)w)w_x] \\ &\leqq (1-\sigma)[w_{xx}+f'(w)w_x] \\ &\leqq 0 \end{aligned} \tag{3.1}$$

since $w_x>0$ on $[0,1]$ and f' is assumed increasing. Moreover on $x=1$,

$$\begin{aligned} v_x-g(v) &= (1-\sigma)g(w)-g((1-\sigma)w) \\ &= (1-\sigma)w[g(w)/w-g((1-\sigma)w)/((1-\sigma)w)] \\ &\geqq 0. \end{aligned} \tag{3.2}$$

We now set

$$\psi(x,t)=e^{(\lambda x+\mu t)}(v(x)-u(x,t)).$$

We find that in $D_{T-\delta}$

$$\psi_t \geqq \psi_{xx}+(-2\lambda+f'(v))\psi_x+[\lambda^2-\mu-\lambda f'(v)+f''(\xi)u_x]\psi$$

and at $x=1$, $0<t\leqq T-\delta$, we have

$$\psi_x \geqq (g'(\eta)+\lambda)\psi.$$

We choose λ, μ to make the coefficients of ψ in these last two inequalities negative. Therefore, if ψ has a negative minimum in $D_{T-\delta}$, it must occur at $x=0$ or at $(1,0)$. At $(1,0)$, however, $\psi(1,0)\geqq 0$. Therefore $\psi\geqq 0$ and the second statement is proved. An argument similar to the above shows us that if $u(x,0)\geqq(1+\sigma)w(x)$, then $u(x,t)\geqq(1+\sigma)w(x)$ on $D_T\cup\Gamma_T$.

We then have, in consequence of the local existence and continuation results, the following theorem.

THEOREM 3.2A. *Let $f+g$ be strictly increasing on $[0,\infty)$. Suppose also that f' is strictly increasing and $g(u)/u$ is increasing on $[0,\infty)$ and that the roots of* (2.2) *are isolated. Then there is at most one positive stationary solution of* (A_1), *call it $w(x)$. Moreover, if $u(\cdot,\cdot)$ solves* (A_1) *on $D_T\cup\Gamma_T$ and $0\leqq u(x,0)\leqq(1-\sigma)w(x)$ on $[0,1]$, then we may take $T=+\infty$ and $0\leqq u(x,t)\leqq(1-\sigma)w(x)$ for all $x\in[0,1]$, $t\in[0,\infty)$. Therefore the null solution is stable from above and $w(x)$ is unstable from below and above (when it exists).*

Proof. Let w_1, w_2 be two stationary solutions of (A_1) with $0<w_1(1)<w_2(1)$ and assume that there are no solutions of (2.2) in $(w_1(1), w_2(1))$. By Theorem 2.1A, we have $w_1(x)<w_2(x)$ on $(0,1]$. Moreover, $w_i'(x)>0$ for $i=1,2$ on $[0,1]$ by Lemma 2.1. With $q(x)=w_1'(x)/w_2'(x)$, we have $q'(x)=(f'(w_2)-f'(w_1))q>0$ on $(0,1]$. From this it follows that $w_1'(0)\leqq w_2'(0)$ which, by uniqueness, must be strict. Moreover,

$$0<\frac{w_1'(1)}{w_2'(1)}=\frac{g(w_1(1))}{g(w_2(1))}=\frac{g(w_1(1))/w_1(1)}{g(w_2(1))/w_2(1)}\cdot\frac{w_1(1)}{w_2(1)}\leqq\frac{w_1(1)}{w_2(1)}<1.$$

Set $\gamma_i=1-w_1'(i)/w_2'(i)$, $i=0,1$. Then $\gamma_i\in(0,1)$ and, on $(0,1]$,

$$(1-\gamma_0)w_2(x)<w_1(x)<(1-\gamma_1)w_2(x).$$

Let $u(x,t)$ solve (A_1) with $u(x,0)=(1-\gamma_1)^{1/2}w_2(x)$. Then

$$(1-\gamma_1)^{-1/2}w_1(x)<u(x,0)\leqq(1-\gamma_1)^{1/2}w_2(x).$$

By the lemma and this inequality,

$$(1-\gamma_1)^{-1/2}w_1(x)\leqq u(x,t)\leqq(1-\gamma_1)^{1/2}w_2(x)$$

on $D_T \cup \Gamma_T$. From this a priori bound and the continuation theorems below, $T = +\infty$. Since $u_t \leqq 0$ on $[0, 1] \times [0, \infty)$ (see Appendix I) $\lim_{t\to\infty} u(x, t) = \phi(x)$ exists and

$$w_1(x) < (1-\gamma_1)^{-1/2} w_1(x) \leqq \phi(x) \leqq (1-\gamma_1)^{1/2} w_2(x) < w_2(x)$$

on $(0, 1]$. Let

$$F(x, t) = \int_0^1 G(x, y) u(y, t)\, dy$$

where

$$G(x, y) = \begin{cases} x & \text{if } 0 \leqq x \leqq y \leqq 1, \\ y & \text{if } 0 \leqq y \leqq x \leqq 1. \end{cases}$$

Then $\lim_{t\to\infty} F(x, t) = \int_0^1 G(x, y)\phi(y)\, dy$ and is finite. If we calculate F_t, we see that

$$\begin{aligned} F_t(x, t) &= \int_0^1 G(x, y) u_t(y, t)\, dy \\ &= -u(x, t) - \int_0^x f(u(y, t))\, dy + x[g(u(1, t)) + f(u(1, t))] \\ &\to -\left[\phi(x) + \int_0^x f(\phi(y))\, dy\right] + x[g(\phi(1)) + f(\phi(1))] \end{aligned}$$

as $t \to \infty$. This limit, which is nonpositive, is in fact zero for $x \in [0, 1]$; otherwise F would not have a finite limit as $t \to +\infty$. Therefore,

$$\phi(x) + \int_0^x f(\phi(y))\, dy = x[g(\phi(1)) + f(\phi(1))]$$

and hence ϕ is a stationary solution of (A_1) with $\phi(1) \in (w_1(1), w_2(1))$, which is the desired contradiction. (If $g(u)/u$ is strictly increasing, then one can relax the condition that the roots of (2.2) are isolated. It then follows that

$$w_1(1)/w_2(1) < 1 - \gamma_1 = w_1'(1)/w_2'(1) < w_1(1)/w_2(1)$$

which is a contradiction and the rest of the argument may be omitted.)

The second statement of the theorem follows from the lemma and the continuation theorems. The null solution is therefore stable from above in the class of continuous functions on $[0, 1]$ vanishing at $x = 0$ while $w(x)$ is unstable from above and below in this class. □

Although the positive stationary solution is unstable (when it exists), there remains the question of the long-time behavior of solutions of (A_1) when $u(x, 0) > w(x)$.

LEMMA 3.3A. *Suppose that* $f'(u) \geqq 0$, $g(u) \geqq 0$ *for* $u \geqq 0$ *and that* $u(x, t)$ *is a nonnegative solution of* (A_1) *on* $D_T \cup \Gamma_T$ *with* $u_x(x, 0) \geqq 0$ *and* $u_x(1, 0) = g(u(1, 0))$. *Then* $u_x(x, t) \geqq 0$ *on* $D_T \cup \Gamma_T$ *and consequently,*

$$u(x, t) \geqq v(x, t)$$

where v *solves*

$$(C_1) \qquad \begin{aligned} &v_t = v_{xx}, && 0 < x < 1, \quad 0 < t < T, \\ &v(0, t) = 0, \\ &v_x(1, t) = g(v(1, t)), && 0 \leqq t < T, \\ &v(x, 0) \leqq u(x, 0), && 0 \leqq x \leqq 1. \end{aligned}$$

Proof. Since $u_x \geqq 0$ on the parabolic boundary and satisfies a linear parabolic equation, $u_x \geqq 0$ in $D_t \cup \Gamma_T$. It follows that $u_t \geqq u_{xx}$ in D_T so that $w = u - v$ satisfies

$w_{xx} - w_t \leqq 0$ in D_T, $w \geqq 0$ when $t = 0$, $w = 0$ on $x = 0$, $w_x = A(x, t)w$ on $x = 1$ where $A(x, t) \equiv (g(u) - g(v))/(u - v)$. By the maximum principle again, $w \geqq 0$. □

LEMMA 3.4A. *Define* $G(u) = \int_0^u g(y)\, dy$ *and suppose that* g *satisfies the structure condition*

$$(p+1)G(u) \leqq ug(u) \tag{3.3}$$

on R^1 *where* $p > 1$ *is given. Let* $v(x, 0) \equiv v_0(x)$. *If* $v_0(x) \geqq 0$, $v_0'(x) \geqq 0$ *and*

$$v_0'(1) = g(v_0(1)), \tag{3.4}$$

$$v_0(0) = 0, \tag{3.5}$$

$$\frac{1}{2}\int_0^1 (v_0'(x))^2\, dx < G(v_0(1)), \tag{3.6}$$

then the solution of (C_1) *blows up in finite time; i.e.,* $T < \infty$ *and*

$$\lim_{t \uparrow T^-} \sup_{0 \leqq x \leqq 1} v(x, t) = +\infty.$$

Proof. The solution (which is nonnegative) fails to be global by the concavity arguments given in [8]. If the solution remains bounded on $\bar{D}_T$, then by the continuation arguments below, it may be continued to $\bar{D}_{T+\delta}$ for some $\delta > 0$. Hence v blows up pointwise in finite time. □

(A variant of this result can be obtained from [13] provided g' exists, an assumption not needed here.)

THEOREM 3.5A. *Let* u *solve* (A_1) *with* $u(x, 0) = v_0(x)$ *and* f, g *as in the preceding lemmas. Then* $u(x, t)$ *blows up in finite time.*

Proof. By Lemmas 3.3A, 3.4A u cannot exist for all time. By the continuation theorems it is continuable if it is bounded on $\bar{D}_T$. Therefore u blows up pointwise in finite time. □

Example 3.1. Suppose $f(u) = \varepsilon u^2/2$, $g(u) = au^p$, $p > 1$. In order to construct such a v_0 as required in Lemma 3.4A, we let

$$v_0(x) = A[(r^2 - (\alpha - x)^2)^{1/2} - (r^2 - \alpha^2)^{1/2}] \tag{3.7}$$

where A, α, r are positive and $r > \alpha > 1$. The constants A, α, r are to be chosen below. Notice that $v_0(0) = 0$, $v_0(x) > 0$ on $(0, 1]$ and $v_0'(x) = A(\alpha - x)(r^2 - (\alpha - x)^2)^{-1/2} > 0$. For any α, r define A by the condition $v_0'(1) = av_0^p(1)$, i.e., by

$$\frac{aA^{p-1}(2\alpha - 1)^p}{[(r^2 - \alpha^2)^{1/2} + (r^2 - (\alpha - 1)^2)^{1/2}]^p} = \frac{\alpha - 1}{(r^2 - (\alpha - 1)^2)^{1/2}}.$$

The final condition of Lemma 3.4A then holds if and only if

$$\frac{a}{p+1} v_0^{p+1}(1) = \frac{(\alpha - 1)(2\alpha - 1) \cdot A^2}{(r^2 - (\alpha - 1)^2)^{1/2}[(r^2 - \alpha^2)^{1/2} + (r^2 - (\alpha - 1)^2)^{1/2}]} \cdot \frac{1}{p+1}$$

$$> \frac{1}{2}\int_0^1 (v_0'(x))^2\, dx = \frac{1}{2}\left\{\frac{1}{2} r \ln\left[\frac{(r - \alpha + 1)(r + \alpha)}{(r + \alpha - 1)(r - \alpha)}\right] - 1\right\} \cdot A^2$$

which will hold if $r > \alpha^2 \gg 1$. One uses the approximation (valid for x small) $\ln(1 + x) \approx x - x^2/2 + x^3/3$ to verify this.

There is a second form of an instability-stability result for (A_1). Somewhat stronger regularity is required however. (See Appendix I.)

THEOREM 3.6A. *In* (A_1) *replace* f *by* εf *where* $\varepsilon \geqq 0$ *is a parameter. Suppose that* $\varepsilon f + g$ *and* f *are* C^1 *increasing functions on* $[0, \infty)$ *with* $f' > 0$ *for* $u > 0$. *Let* $w(x, \varepsilon)$ *be a* C^1 *(in* ε*) branch of positive solutions on some* ε *interval and let* $w_1(\varepsilon) \equiv w(1, \varepsilon)$. *If* $w_1'(\varepsilon) > 0$ *on this branch, the solutions are stable while if* $w_1'(\varepsilon) < 0$ *on this branch, the solutions are unstable. (Here* $w_1'(\varepsilon) \equiv \partial w(1, \varepsilon)/\partial \varepsilon$.)

Proof. We know that

$$x = \int_0^{w(x,\varepsilon)} \frac{d\sigma}{g(w_1(\varepsilon)) + \varepsilon(f(w_1(\varepsilon)) - f(\sigma))} \tag{3.8}$$

on $[0, 1]$. Differentiating with respect to ε, we obtain

$$\begin{aligned} &\{g(w_1(\varepsilon)) + \varepsilon[f(w_1(\varepsilon)) - f(w(x, \varepsilon))]\}^{-1} \frac{\partial w}{\partial \varepsilon}(x, \varepsilon) \\ &= \int_0^{w(x,\varepsilon)} \{[g'(w_1) + \varepsilon f'(w_1)]w_1'(\varepsilon) + (f(w_1) - f(\sigma))\} \frac{d\sigma}{D^2(\sigma)} \end{aligned} \tag{3.9}$$

where $D(\sigma)$ is the denominator in (3.8). Thus, if $w_1'(\varepsilon) > 0$, $\partial w(x, \varepsilon)/\partial \varepsilon > 0$ on $(0, 1]$. Therefore $w(x, \varepsilon_1) < w(x, \varepsilon_2)$ on $[0, 1)$ if $[\varepsilon_1, \varepsilon_2]$ is contained in the domain of this branch. Also $w_x(x, \varepsilon_i) > 0$ on $[0, 1]$.

Suppose that $u(x, t, \varepsilon_1)$ is a solution of (A_1) with $u(x, 0, \varepsilon_1) = w(x, \varepsilon_2)$. Then, on $(0, 1)$,

$$\begin{aligned} u_t(x, 0, \varepsilon_1) &= w_{xx}(x, \varepsilon_2) + \varepsilon_1 f'(w(x, \varepsilon_2)) \cdot w_x(x, \varepsilon_2) \\ &< w_{xx}(x, \varepsilon_2) + \varepsilon_2 f'(w(x, \varepsilon_2)) \cdot w_x(x, \varepsilon_2) \\ &= 0. \end{aligned}$$

Therefore since u_t satisfies a linear problem with homogeneous boundary data, $u_t < 0$ on $D_T \cup \Gamma_T$ except at $x = 0$ and

$$w(x, \varepsilon_1) < u(x, t, \varepsilon_1) < w(x, \varepsilon_2). \tag{3.10}$$

(The first inequality follows from standard comparison theorems. Note that u and the w's satisfy the same boundary conditions.) Thus, from the continuation theorems and (3.10), $T = \infty$ and

$$\phi(x, \varepsilon_1) \equiv \lim_{t \to \infty} u(x, t, \varepsilon_1)$$

exists. From (3.10), $w(x, \varepsilon_1) \leqq \phi(x, \varepsilon_1) < w(x, \varepsilon_2)$ so that, letting $\varepsilon_2 \downarrow \varepsilon_1$, we obtain $\phi(x, \varepsilon_1) = w(x, \varepsilon_1)$. This suffices to show that $w(x, \varepsilon_1)$ is stable from above. Similarly, with $\varepsilon_1 > \varepsilon_2$, one easily shows that $w(x, \varepsilon_1)$ is stable from below.

In the second case, from $w_1'(\varepsilon) < 0$ on $[\varepsilon_1, \varepsilon_2]$ we have that $w(1, \varepsilon_2) < w(1, \varepsilon_1)$ and consequently $w(x, \varepsilon_2) < w(x, \varepsilon_1)$ is a left open neighborhood of $x = 1$.

Suppose that $u(x, t, \varepsilon_2)$ is a solution of (A_1) with $u(x, 0, \varepsilon_2) = w(x, \varepsilon_1)$. Then since $f' > 0$, $w_x \geqq 0$, on $(0, 1)$,

$$\begin{aligned} u_t(x, 0, \varepsilon_2) &= w_{xx}(x, \varepsilon_1) + \varepsilon_2 f'(w(x, \varepsilon_1))w_x(x, \varepsilon_1) \\ &> w_{xx}(x, \varepsilon_1) + \varepsilon_1 f'(w(x, \varepsilon_1))w_x(x, \varepsilon_1) = 0. \end{aligned}$$

Since u_t is nonnegative at $x = 0$ and satisfies a homogeneous linear condition at $x = 1$, $u_t > 0$ in D_T. Therefore u is increasing in t. Hence $w(x, \varepsilon_2)$ is unstable from above. A similar argument with $\varepsilon_1 > \varepsilon_2$ shows that $w(x, \varepsilon_2)$ is unstable from below. □

We have the following corollary of Theorem 3.6A.

COROLLARY 3.7A. *Let f, g be as in the preceding theorem. If $g'(w_1(0))<1$, the branch of stationary solutions emanating from $\varepsilon=0$ is stable while if $g'(w_1(0))>1$, it is unstable.*

Proof. We need to compute $w_1'(\varepsilon)$ on such branches. Setting $x=1$ in (3.9), we find that

$$\left[1-g(w_1)(g'(w_1)+\varepsilon f'(w_1))\int_0^{w_1(\varepsilon)}\frac{d\sigma}{D^2(\sigma)}\right]w_1'(\varepsilon) = g(w_1)\int_0^{w_1(\varepsilon)}\frac{[f(w_1)-f(\sigma)]}{D^2(\sigma)}\,d\sigma \tag{3.11}$$

where $D(\sigma)$ denotes the denominator in (3.8). When $\varepsilon=0$, $g(w_1(0))=w_1(0)$. We find from (3.11) that

$$[1-g'(w_1(0))]w_1'(0)=\frac{1}{w_1(0)}\int_0^{w_1(0)}[f(w_1(0))-f(\sigma)]\,d\sigma. \tag{3.12}$$

Therefore $w_1'(0)<0$ or $w_1'(0)>0$, accordingly $g'(w_1(0))>1$ or $g'(w_1(0))<1$. From (3.11), it follows that the sign of $w_1'(\varepsilon)$ cannot change along the branch unless the coefficient of $w_1'(\varepsilon)$ changes sign. Since the product is strictly positive, $w_1'(\varepsilon)$ will be of constant sign. □

As an example, with $f(u)=\frac{1}{2}\varepsilon u^2$, $g(u)=au^p$, we find that $w_1(0)=a^{-1/(p-1)}$ and

$$1-g'(w_1(0))=1-p. \tag{3.13}$$

Thus, positive solutions of (A) are unstable if $p>1$. From Theorem 3.2A, the zero solution is stable. For $0<p<1$, on the branch emanating from $(0, w_1(0))$, we have $w_1'(\varepsilon)>0$. Therefore stationary solutions are stable on this branch. For $w_1'(\varepsilon)<0$ (the upper branch in Figs. 1.1 and 1.2 with $0<p<1$) the stationary solutions are unstable.

There remains only the question of the stability of the positive solution when $p=1$ and the stability of the null solution when $0<p\leqq 1$. When $p=1$ and $0<a<1$, we see from Example 2.1 that $w_1(\varepsilon)=2a/(\varepsilon v_1(a))$. Thus $w_1'(\varepsilon)<0$ and the branch of positive solutions is unstable. Therefore, from Theorem 3.2A, the null solution is stable.

In order to demonstrate the instability of the null solution when $p=1$ and $a>1$ or when $0<p<1$, suppose $u_0(x)>0$ on $(0,1]$ and $u_0(0)=0$, $u_0'(1)\geqq au_0^p(1)$. Choose $\lambda>0$, $\delta>0$ so small that

$$a\geqq \delta^{1-p}\lambda\coth(\lambda)(\sinh(\lambda))^{1-p},$$

$$u_0(x)\geqq \delta\sinh(\lambda x).$$

Then

$$v(x,t)=\delta\sinh(\lambda x) \tag{3.14}$$

is a subsolution. That is, $v_{xx}+\varepsilon vv_x\geqq 0$, $v_x(1,t)\leqq a(v(1,t))^p$. Therefore $u(x,t)\geqq \delta\sinh(\lambda x)$ for all t in the existence interval and hence zero is unstable from above. (When $a=1$, use $v=\delta x$.)

We next turn our attention to (B_1). As noted earlier, the structure conditions on f, g are somewhat different in this problem. There are parallel results however.

LEMMA 3.1B. *Let $f'(u)$ be increasing. Let $g(u)/u$ be bounded in a neighborhood of $u=0$ or $u>0$ on $\{0\}\times[0,T)$. Let u be a solution of (B_1) on $D_T\cup\Gamma_T$. If $u(x,0)>0$ on $[0,1)$, then $u>0$ on $D_T\cup\Gamma_T$ except at $x=1$. Suppose that $g(u)/u$ is decreasing on $(0,\infty)$. Suppose that $w(x)$ is a positive stationary solution of (B_1) and $u(x,0)\leqq(1+\sigma)w(x)$ on $[0,1]$, then $u(x,t)\leqq(1+\sigma)w(x)$ on $D_T\cup\Gamma_T$, while if $u(x,0)\geqq(1-\sigma)w(x)$ on $[0,1]$, then $u(x,t)\geqq(1-\sigma)w(x)$ on $D_T\cup\Gamma_T$.*

Proof. The proof of the lemma proceeds (mutatis mundanis) in the same manner as that of Lemma 3.1A. We note that this time $w'(x)<0$ on $[0,1]$. Therefore, with $v(x)=(1-\sigma)w(x)$ $(0<\sigma<1)$ we have

$$\begin{aligned} v_{xx}+f'(v)v_x &= (1-\sigma)[w_{xx}+f'((1-\sigma)w)w_x] \\ &= (1-\sigma)w_x[f'((1-\sigma)w)-f'(w)] \\ &\geqq 0. \end{aligned} \tag{3.15}$$

While at $x=0$,

$$\begin{aligned} v_x+g(v) &= g((1-\sigma)w)-(1-\sigma)g(w) \\ &= (1-\sigma)w[g((1-\sigma)w)/(1-\sigma)w-g(w)/w] \\ &\leqq 0. \end{aligned} \tag{3.16}$$

The inequalities (3.15) and (3.16) are reversed for $v(x)$ when $v(x)=(1+\sigma)w(x)$. □

THEOREM 3.2B. *Let f', $-g(u)/u$ be increasing on $[0,\infty)$ with f', g strictly increasing. Suppose that the roots of* (2.4) *are isolated and satisfy the conditions of Theorem* 2.1B. *Then there is at most one positive stationary solution $w(x)$ of* (B_1) *and it is stable. The null solution is unstable from above.*

Proof. The stability and instability follow from Lemma 3.1B and the continuation theorems below. Let w_1, w_2 be two positive stationary solutions and suppose $w_1(x)<w_2(x)$ on $[0,1)$ and that there are no roots of (2.4) in $(w_1(0), w_2(0))$. It follows from Lemma 2.1 that $w_i'(x)<0$ on $[0,1]$. Exactly as in the proof of Theorem 3.2A, there is a constant $\gamma_0=1-w_1'(0)/w_2'(0)$ in $(0,1)$ such that

$$w_1(x)<(1-\gamma_0)w_2(x).$$

(If $\gamma_0=0$, then $w_1'(0)=w_2'(0)$ or $g(w_1(0))=g(w_2(0))$. Then $w_1(0)=w_2(0)$ and $w_1\equiv w_2$. If $\gamma_0=1$, then $w_1'(0)=0=g(w_1(0))=0$ and hence $w_1(0)=0$. Then $w_1\equiv 0$.) Let $\delta>0$ such that

$$(1-\gamma_0)w_2(0)<(1+\delta)w_1(0)<w_2(0).$$

If we set

$$\begin{aligned} v(x) &= ((1+\delta)/(1-\gamma_0))w_1(x)-w_2(x) \\ &\equiv (1+\bar{\delta})w_1(x)-w_2(x), \end{aligned}$$

then $v(0)>0$, $v(1)=0$ and

$$\begin{aligned} v''(x) &= -(1+\delta)f'(w_1)w_1'+f'(w_2)w_2' \\ &< -f'(w_2)v'(x). \end{aligned}$$

Consequently v cannot have a minimum on $(0,1)$ and $v(x)>0$ on $[0,1)$. Therefore

$$(1+\delta)w_1(x)>(1-\gamma_0)w_2(x).$$

If we let u solve (B_1) with $u(x,0)=(1+\delta)w_1(x)$, then, by the lemma $u(x,t)>0$ except at $x=1$ on $D_T\cup\Gamma_T$ and by the first and second maximum principles

$$u(x,t)<(1+\delta)w_1(x)$$

on D_T and Γ_T. The lemma assures us that for all x, t

$$u(x,t)>(1-\gamma_0)w_2(x).$$

Therefore, by the continuation theorems, we may take $T=+\infty$. However, by the lemma $u(x,t)\leqq u(x,0)$ on $[0,1]$ so that $u_t\leqq 0$ for all t (see Appendix I) and hence $\lim_{t\to\infty} u(x,t)=\phi(x)$ exists for all $x\in[0,1]$. Exactly as in Theorem 3.2A, we easily establish that ϕ is a stationary solution and hence $\phi(0)$ is a root of (2.4). Since

$$w_1(0)<(1-\gamma_0)w_2(0)<\phi(0)<(1+\delta)w_1(0)<w_2(0),$$

we have reached the desired contradiction. □

As in Theorem 3.2A, if $-g(u)/u$ is strictly increasing, the proof may be shortened and the zeros of (2.4) need not be assumed to be isolated.

The choice $f(u)=\frac{1}{2}\varepsilon u^2$, $g(u)=au^p$, $a>0$, $0<p<1$, together with the observations in Example 2.1, provides an illustration of this result.

Sometimes solutions of (B_1) can blow up in finite time.

LEMMA 3.3B. *Suppose that $g(u)\geqq 0$ and that u solves (B_1) on $D_T\cup\Gamma_T$ with $u(x,0)>0$ on $[0,1)$. Then $u(x,t)>0$ on $D_T\cup\Gamma_T$ except at $x=1$. If $u_x(x,0)\leqq 0$, then $u_x(x,t)\leqq 0$ while if $u_t(x,0)\geqq 0$ then $u_t(x,t)\geqq 0$ on $D_T\cup\Gamma_T$.*

The proof of this rests on straightforward applications of the maximum principle and is omitted.

LEMMA 3.4B. *Let $f'\geqq 0$ and define $f_1(u)=\int_0^u f(\eta)\,d\eta$, $G(u)=\int_0^u g(\eta)\,d\eta$. Suppose that u solves (B_1) on $D_T\cup\Gamma_T$ and that the hypotheses of Lemma 3.3B hold. Define, for $t\leqq T$,*

$$F(t)=\int_0^t\int_0^1 u^2(x,\eta)\,dx\,d\eta+(T^*-t)\int_0^1 u^2(x,0)\,dx+\beta(t+t_0)^2 \tag{3.17}$$

where t_0, β, T^* are positive constants with $T^*\geqq T$. Then, for any $\alpha>0$, on $[0,T)$, we have,

$$\begin{aligned} FF''-(\alpha+1)(F')^2\geqq 4(\alpha+1)F\Big[G(u(0,0))-\frac{1}{2}\int_0^1 u_x^2(x,0)\,dx \\ -(2\alpha+1)\beta/(2\alpha+2)\Big]+2FQ(u(0,t)) \end{aligned} \tag{3.18}$$

where

$$Q(v)=vg(v)-2(\alpha+1)G(v)+f_1(v)-vf(v).$$

Proof. The proof is a straightforward calculation, variants of which can be found in [7], for example (in the case $f=0$). We find

$$\begin{aligned} FF''-(\alpha+1)(F')^2&=4(\alpha+1)S^2+2FQ(u(0,t)) \\ &\quad+4(\alpha+1)F\Big[G(u(0,0))-\frac{1}{2}\int_0^1 u_x^2(x,0)\,dx \\ &\quad-(2\alpha+1)\beta/(2\alpha+2)\Big] \\ &\quad-4(\alpha+1)F\cdot\int_0^t\int_0^1 u_xu_\eta f'(u)\,dx\,d\eta \end{aligned} \tag{3.19}$$

where

$$\begin{aligned} S^2=&\Big(\int_0^t\int_0^1 u^2\,dx\,d\eta+\beta(t+t_0)^2\Big)\Big(\int_1^t\int_0^1 u_\eta^2\,dx\,d\eta+\beta\Big) \\ &-\Big(\int_0^t\int_0^1 uu_\eta\,dx\,d\eta+\beta(t+t_0)\Big)^2. \end{aligned}$$

□

In view of the preceding lemma, we see that the double integral on the right of (3.19) is nonpositive and (3.18) follows immediately.

THEOREM 3.5B. *Let u solve* (B_1) *with f, g as in the above lemmas. Assume that*

$$u(x,0)>0 \quad \text{on } [0,1), \quad u(1,0)=0, \tag{3.20}$$

$$u_t(x,0)\geqq 0 \quad \text{on } (0,1), \tag{3.21}$$

$$u_x(x,0)\leqq 0 \quad \text{on } (0,1), \tag{3.22}$$

$$\frac{1}{2}\int_0^1 u_x^2(x,0)\,dx < G(u(0,0)), \tag{3.23}$$

$$\text{There is } \alpha>0 \text{ such that } Q(v)\geqq 0, \tag{3.24}$$

$$-u_x(0,0)=g(u(0,0)). \tag{3.25}$$

Then u cannot exist for all time.

Proof. If we can establish the theorem, the continuation theorems below permit the conclusion that u blows up in finite time. By choosing β such that

$$0<\beta\leqq\frac{2(\alpha+1)}{2\alpha+1}\cdot\left[G(u(0,0))-\frac{1}{2}\int_0^1 u_x^2(x,0)\,dx\right],$$

and noting (3.24) we see that $FF''-(\alpha+1)(F')^2\geqq 0$ on $[0,T)$ and therefore $(F^{-\alpha})''\leqq 0$ there. If $T=\infty$, T^* will be at our disposal. We find that $F^{-\alpha}$ has a zero at some value $\bar{t}\leqq F(0)/\alpha F'(0)$, provided $F(0)/\alpha F'(0)$ does not exceed T^*. These conditions are satisfied if we take $t_0=\alpha T^*$ and

$$T^*=(\alpha^2\beta)^{-1}\int_0^1 u^2(x,0)\,dx$$

while then $\bar{t}\leqq T^*$. □

Example 3.2. Let $f(u)=\frac{1}{2}\varepsilon u^2$, $g(u)=au^p$ and

$$u(x,0)=\frac{A[(\gamma-x)^2-(\gamma-1)^2]}{\sqrt{r^2-(\gamma-1)^2}+\sqrt{r^2-(\gamma-x)^2}}$$

where $A>0$, $r>\gamma>1$. We shall choose A, γ, r so that (3.20)-(3.25) hold. Clearly $u(1,0)=0$ and $u(x,0)>0$ on $[0,1)$. Also $u_x(x,0)=-A(\gamma-x)/(r^2-(\gamma-x)^2)^{1/2}<0$. We choose A so that (3.25) holds, i.e.,

$$A=\left[\frac{\gamma}{a(2\gamma-1)^p\sqrt{r^2-\gamma^2}}(\sqrt{r^2-\gamma^2}+\sqrt{r^2-(\gamma-1)^2})^p\right]^{1/(p-1)}.$$

If r, $\gamma\gg 1$ and $r\gg\gamma$, then

$$A\approx a^{-1/(p-1)}(r^2-\gamma^2)^{1/2}/\gamma.$$

In order to check that $u_t(x,0)\geqq 0$, we compute

$$\begin{aligned}W(x)&\equiv u_{xx}(x,0)+\varepsilon u(x,0)u_x(x,0)\\&=\frac{A}{\sqrt{r^2-(\gamma-x)^2}}\left[\frac{r^2}{(r^2-(\gamma-x)^2)}-\frac{\varepsilon A(2\gamma-x-1)(1-x)(\gamma-x)}{\sqrt{r^2-(\gamma-1)^2}+\sqrt{r^2-(\gamma-x)^2}}\right].\end{aligned}$$

Since $(\gamma-1)^2\leqq(\gamma-x)^2<\gamma^2$ on $[0,1]$, we find that

$$W(x)\geqq\frac{A}{(r^2-(\gamma-x)^2)^{3/2}}[r^2-\varepsilon\gamma^2\sqrt{r^2-(\gamma-1)^2}].$$

For $r, \gamma \gg 1$ and $r \gg \gamma$, the factor in brackets is larger than $\frac{1}{2}r^2$. Conditions (3.20), (3.21), (3.22) and (3.25) thus hold.

Condition (3.23) is verified exactly as in Example 3.1. We note that

$$\frac{a(u(0,0))^{p+1}}{p+1} = \frac{-u_x(0,0)u(0,0)}{(p+1)} \approx \frac{A^2\gamma^2}{(r^2-\gamma^2)(p+1)}$$

while

$$\frac{1}{2}\int_0^1 u_x^2(x,0)\,dx = \frac{A^2}{2}\left\{-1+\frac{r}{2}\ln\left[\frac{(r-\gamma+1)(r+\gamma)}{(r+\gamma-1)(r-\gamma)}\right]\right\}.$$

In order to check (3.24) we see that for $u>0$,

$$Q(u) = a(1-2(\alpha+1)/(p+1))u^{p+1} - \tfrac{1}{3}\varepsilon u^3.$$

For $p=2$, $Q(u)\geqq 0$ for $u\geqq 0$ provided $a>\varepsilon$. For $p>2$, $Q(u)$ is positive to the right of

$$\bar{u} = \left(\frac{\varepsilon}{3}\frac{(p+1)}{(p-1)}\right)^{1/(p-2)} a^{-1/(p-2)}.$$

Since $u_t(0,t)\geqq 0$, $u(0,t)\geqq \bar{u}$ and $Q(u)\geqq 0$ provided

$$u(0,0)\geqq \bar{u}$$

which in turn holds for $r, \gamma \gg 1$, $0<r\gg\gamma$ if

$$a > \left(\frac{\varepsilon(p+1)}{3(p-1)}\right)^{p-1}.$$

The question of finite time blow up remains open if $\varepsilon \gg a$.

We next replace f by εf in (B_1). We prove the following.

THEOREM 3.6B. *Let $f'>0$ on $(0,\infty)$ and suppose that $w(x,\varepsilon)$ is a C^1 branch of positive stationary solutions of (B_1) with f replaced by εf along which $g(w_0)-\varepsilon f(w_0)>0$ $(w_0(\varepsilon)\equiv w(0,\varepsilon))$. If $w_0'(\varepsilon)=\partial w(0,\varepsilon)/\partial\varepsilon>0$, this is a branch of unstable stationary solutions. If $f''\geqq 0$ and $w_0'(\varepsilon)<0$, this is a branch of stable stationary solutions.*

Proof. Suppose $w_0'(\varepsilon)>0$. Then if $\varepsilon_1<\varepsilon_2$, $w(x,\varepsilon_1)<w(x,\varepsilon_2)$ in a neighborhood of $x=0$. Then, with $u(x,t,\varepsilon_1)$ a solution of (B_1) with $\varepsilon=\varepsilon_1$ such that $u(x,0,\varepsilon_1)=w(x,\varepsilon_2)$, we find that on $(0,1)$,

$$\begin{aligned} u_t(x,0,\varepsilon_2) &= w''+\varepsilon_1 f'(w)w' \\ &= w''+\varepsilon_2 f'(w)w'+(\varepsilon_1-\varepsilon_2)f'(w)w' \\ &> 0 \end{aligned}$$

since $w'<0$, $f'>0$ and $\varepsilon_1-\varepsilon_2<0$. From this we conclude as before, that $u_t(x,t)>0$ on $D_T\cup\Gamma_T$ and hence that $w(x,\varepsilon_1)$ is unstable from above. A similar argument with $u(x,0,\varepsilon_2)=w(x,\varepsilon_1)$ shows that $w(x,\varepsilon_2)$ is unstable from below.

Suppose next that $w_0'(\varepsilon)<0$. Then $v(x)\equiv\partial w(x,\varepsilon)/\partial\varepsilon$ solves, on $(0,1)$,

$$v''+\varepsilon f'(w)v'+\varepsilon w'f''(w)v = -f'(w)w' > 0,$$

since $f'(w)>0$ if $w>0$ and $w'<0$ on $[0,1]$. Therefore, since $v(0)<0$, $v(1)=0$, $v(x)<0$ on $[0,1)$. It follows that if $0<\varepsilon_1<\varepsilon_2$ on the branch, then $w(x,\varepsilon_1)>w(x,\varepsilon_2)$ on $[0,1)$. Again, with $u(x,0,\varepsilon_1)=w(x,\varepsilon_2)$, we find that

$$\begin{aligned} u_t(x,0,\varepsilon_1) &= w''+\varepsilon_1 f'(w)w' \\ &= w''+\varepsilon_2 f'(w)w'-(\varepsilon_2-\varepsilon_1)f'(w)w' \\ &> 0 \end{aligned}$$

and consequently on $D_T \cup \Gamma_T$, by this and comparison

$$w(x, \varepsilon_2) \leqq u(x, t, \varepsilon_1) \leqq w(x, \varepsilon_1).$$

(Again note that u and the w's satisfy the same boundary conditions.) Thus, we may take $T = +\infty$ and $\lim_{t\to\infty} u(x, t, \varepsilon_1) = \phi(x, \varepsilon_1)$ exists. Moreover, $w(x, \varepsilon_2) \leqq \phi(x, \varepsilon_1) \leqq w(x, \varepsilon_1)$. Letting ε_2 decrease to ε_1, $\phi(x, \varepsilon_1) = w(x, \varepsilon_1)$. This shows that $w(x, \varepsilon_1)$ is stable from below. A similar argument shows that $w(x, \varepsilon_2)$ is stable from above. □

From (2.4), in this case, we have

$$1 = \int_0^{w_0(\varepsilon)} \frac{d\sigma}{g(w_0(\varepsilon)) + \varepsilon(f(\sigma) - f(w_0(\varepsilon)))}. \tag{3.26}$$

Differentiating with respect to ε along a branch of solutions of (2.4), we find

$$\left\{1 - [g'(w_0) - \varepsilon f'(w_0)]g(w_0) \int_0^{w_0(\varepsilon)} \frac{d\sigma}{D^2(\sigma)}\right\} w_0'(\varepsilon) = g(w_0) \int_0^{w_0(\varepsilon)} \frac{[f(\sigma) - f(w_0)]}{D^2(\sigma)}\, d\sigma$$

where $D(\sigma)$ is the denominator in (3.26). If the branch is defined on an interval $[0, \varepsilon_0)$, we have, at $\varepsilon = 0$,

$$[1 - g'(w_0(0))]w_0'(0) = -\frac{1}{w_0(0)} \int_0^{w_0(0)} [f(w_0(0)) - f(\sigma)]\, d\sigma$$

because again $g(w_0(0)) = w_0(0)$.

Thus, the corollary dual to Corollary 3.7A is the following.

COROLLARY 3.7B. *Let f, g be as in the preceding theorem. If $g'(w_1(0)) > 1$, the branch of stationary solutions emanating from $\varepsilon = 0$ is unstable, while if $g'(w_1(0)) < 1$, this branch is stable.*

The convexity of f in Theorem 3.6B is not necessary.

COROLLARY 3.8B. *Let $f' > 0$ on $(0, \infty)$, and suppose that $w_0'(\varepsilon) < 0$ on some C^1 branch $w(x, \varepsilon)$ of stationary solutions of (B_1) with f replaced by εf. If $g'(w_0(\varepsilon)) - \varepsilon f'(w_0(\varepsilon)) > 0$ along this branch, then the branch consists only of stable stationary solutions.*

Proof. It suffices to show that $\partial w(x, \varepsilon)/\partial\varepsilon < 0$. However, we must have

$$1 - x = \int_0^{w(x,\varepsilon)} \frac{d\sigma}{g(w_0) - \varepsilon f(w_0) + \varepsilon f(\sigma)} \tag{3.27}$$

along such a C^1 branch of stationary solutions and therefore

$$[g(w_0(\varepsilon)) - \varepsilon f(w_0(\varepsilon)) + \varepsilon f(w(x, \varepsilon))]^{-1} \frac{\partial w(x, \varepsilon)}{\partial \varepsilon}$$

$$= \int_0^{w(x,\varepsilon)} \{[g'(w_0) - \varepsilon f'(w_0)]w_0'(\varepsilon) + [f(\sigma) - f(w_0)]\} \frac{d\sigma}{D^2(\sigma)}$$

where $D(\sigma)$ is the denominator in (3.27). By hypothesis, the integrand in the integral is negative while the coefficient of $\partial w/\partial\varepsilon$ is positive. □

As an example, we again take $\varepsilon f(u) = \frac{\varepsilon}{2}u^2$, $g(u) = au^p$, $u \geqq 0$. We find again that $w_0(0) = a^{-1/(p-1)}$ and

$$1 - g'(w_0(0)) = 1 - p.$$

Thus, if $p > 1$, the branch of positive solutions emanating from zero is unstable. For $p > 2$, this branch exists for all $\varepsilon > 0$. Therefore, for all $\varepsilon > 0$, the null solution is stable from above. For $1 < p < 2$, the upper branch is stable since $w_0'(\varepsilon) < 0$ there. When $0 < p < 1$, $w_0'(\varepsilon) < 0$ and the branch of positive solutions (which also exists for all $\varepsilon > 0$) is stable. Thus, by Theorem 3.2B, in this case the null solution is unstable from above.

There remains only the question of the stability of the positive solution when $p=1$ and $a \geqq 1$ and of the null solution when $1 \leqq p \leqq 2$. (For $1<p<2$ we have this stability for $\varepsilon < \varepsilon(p)$.) When $p=1$ and $a \geqq 1$, we find from Example 2.1 that

$$w_0(\varepsilon) = 2a(\alpha_J^2+1)/\varepsilon$$

where α_J is the unique root of $J(\alpha)$. Therefore $w_0'(\varepsilon)<0$ and this is a branch of stable stationary solutions. Again by Theorem 3.2B, it follows that the null solution is unstable in this case.

When $p=1$ and $a<1$ or when $p>1$, set, for $\lambda>0$, $\delta>0$

$$v(x, t) = \delta \sin((1-x)\lambda)\, e^{-\lambda^2 t}.$$

Then one easily checks that $v_{xx}+\varepsilon v v_x \leqq v_t$ and that $v_x(0, t) \leqq -av^p(0, t)$ provided that $\lambda, \delta > 0$ satisfy

$$a \leqq \lambda \cot \lambda \cdot (\delta \sin \lambda)^{-(p-1)}.$$

If $u_0(x) \leqq \delta \sin((1-x)\lambda)$, it then follows that $u(x, t) \leqq v(x, t)$ and the null solution is thus stable from above (indeed asymptotically stable from above).

Finally, it is perhaps worth noting that for both (A) and (B) when $a=0$, there are no nontrivial stationary solutions. It is not difficult to show that the null solution is asymptotically stable.

4. Local existence and continuation. In this section we shall establish the existence of solutions of (A_1) and (B_1) on $D_T \cup \Gamma_T$ for sufficiently small T and certain initial values. This result follows from results in [2]. However, we include an elementary proof here for completeness.

We assume that f, g are defined on R^1, that $g(u)>0$ for $u>0$ and that $f(0)=g(0)=0$. We shall also assume that f is uniformly Lipschitz in compact subsets of R^1, that g is continuous and is uniformly Lipschitz on compact subsets of $R^1-\{0\}$. (This last restriction is necessary to include functions such as $g(u)=|u|^{p-1}u$ where $0<p<1$.) We shall also define

$$f_M \equiv \sup_{|u| \leqq M} |f(u)| \tag{4.1}$$

and

$$g_M \equiv \sup_{|u| \leqq M} |g(u)|. \tag{4.2}$$

We shall discuss problem (A_1). The arguments for (B_1) are similar and are omitted.

Let $G(x, y; t)$ denote Green's function for

$$\mathbf{L}u = u_t - u_{xx}, \qquad \sigma < x < 1, \quad t>0$$

with boundary conditions

$$u(0, t) = u_x(1, t) = 0, \qquad t>0,$$

i.e.,

$$G(x, y; t) = 2\sum_{n=1}^{\infty} \sin(\lambda_n x) \sin(\lambda_n y)\, e^{-\lambda_n^2 t}$$

where $\lambda_n = \frac{1}{2}(2n-1)\pi$. Then $G_x(1, y; t) = G_y(x, 1; t) = G(0, y; t) = G(x, 0; t) = 0$ and u is a solution of (A_1) on $[0, 1] = [0, T)$ if and only if for (x, t) $[0, 1] \times [0, T)$,

$$u(x,t) = \int_0^1 G(x, y; t) u_0(y)\, dy - \int_0^t \int_0^1 G_y(x, y; t-\eta) f(u(y, \eta))\, dy\, d\eta$$

$$+ \int_0^t G(x, 1; t-\eta)[f(u(1, \eta)) + g(u(1, \eta))]\, d\eta \tag{4.3}$$

$$\equiv \mathbf{T}u(x, t).$$

In order to show that (4.3) is solvable for sufficiently small T, we use a contraction mapping argument. We define

$$u_1(x, t) = 0 \tag{4.4}_1$$

and then, iteratively define

$$u_{n+1}(x, t) = \mathbf{T}u_n(x, t). \tag{4.4}_2$$

THEOREM 4.1. *Let the initial datum for problem* (A_1) *be continuous on* $[0, 1]$ *and satisfy*

$$0 < d_1 < \int_0^1 G(1, y; t) u_0(y)\, dy \tag{4.5}$$

for $0 \leqq t \leqq 1$, *say. Then for sufficiently small* T, (A_1) *has a unique solution which satisfies*

$$u(1, t) \geqq d_1/2 \tag{4.6}$$

on $[0, T]$. *The solution is* C^1 *in* t *and* c^2 *in* x *on* $(0, 1) \times (0, T)$ *and continuous on* $\bar{D}_T$. (*If* g *is uniformly Lipschitz continuous on compact subsets of* R^1, (4.5) *and* (4.6) *may be omitted.*)

Proof. The proof is fairly standard. We shall only sketch the arguments.

First, define

$$d_2 = \sup_{0 \leqq x \leqq 1} |u_0(x)|,$$

$$\mu(t) = \sup_{\substack{0 \leqq x \leqq 1 \\ 0 \leqq t' \leqq t}} \int_0^{t'} G(x, 1; t' - \eta)\, d\eta,$$

$$\nu(t) = \sup_{\substack{0 \leqq x \leqq 1 \\ 0 \leqq t' \leqq t}} \int_0^{t'} \int_0^1 |G_y(x, y; t' - \eta)|\, dy\, d\eta.$$

Clearly $\mu(t) \to 0$ monotonically as $t \to 0^+$. Inspection of the principle part of G shows us that the same is true for $\nu(t)$. For fixed $M > d_2$, choose T so small that $T \leqq 1$ and

$$\nu(T) f_M + \mu(T)(f_M + g_M) < \max(M - d_2, \tfrac{1}{2} d_1), \tag{4.7}$$

$$\beta_1 \equiv \nu(T) \cdot \sup_{|\xi| \leqq M} |f'(\xi)| < 1, \tag{4.8}$$

$$\beta_2 \equiv \mu(T)\left(\sup_{|\xi| \leqq M} |f'(\xi)| + \sup_{\substack{d_1/2 \leqq \xi_1 < \xi_2 \leqq M \\ -M \leqq \xi_2 < \xi_1 < -d_1/2}} |g(\xi_2) - g(\xi_1)| / |\xi_2 - \xi_1| \right) < 1. \tag{4.9}$$

It then follows from $(4.4)_1$, $(4.4)_2$, (4.7) and induction that on $\bar{D}_T$

$$\|u_n\|_{\bar{D}_T} \equiv \sup_{\bar{D}_T} |u_n(x, t)| \leqq M \tag{4.10}$$

for all $n = 1, 2, \cdots$. Moreover, on $x = 1$, we have from $(4.4)_2$ and (4.7),

$$u_n(1, t) \geqq \tfrac{1}{2} d_1 \tag{4.11}$$

for $n = 2, 3, \cdots$. A standard estimate then shows us that

$$\begin{aligned} \delta_n &\equiv \| u_{n+1}(1, \cdot) - u_n(1, \cdot) \|_{[0,T]} \\ &\equiv \sup_{0 \leqq t \leqq T} |u_{n+1}(1, t) - u_n(1, t)| < \beta_2 \delta_{n-1} \end{aligned} \tag{4.12}$$

and therefore

$$\delta_n < \beta_2^{n-1} \delta_1 . \tag{4.13}$$

From this it follows that the boundary values $u_n(1, t)$ are uniformly convergent. (From (4.4) $u_n(0, t) \equiv 0$.)

Again, a standard argument shows that if

$$\gamma_n \equiv \| u_{n+1} - u_n \|_{\bar{D}_T}, \tag{4.14}$$

then

$$\gamma_{n+1} \leqq \beta_1 \gamma_n + \delta_n . \tag{4.15}$$

Letting

$$\beta_3 = \max (\beta_1, \beta_2) < 1$$

we obtain from (4.13), (4.15) and the discrete version of Gronwall's inequality

$$\gamma_{n+1} \leqq \gamma_1 \beta_1^n + n \delta_1 \beta_3^{n-1} . \tag{4.16}$$

Therefore $\{u_n\}$ is uniformly convergent on $\bar{D}_T$ and

$$u(x, t) = \lim_{n \to \infty} u_n(x, t) \tag{4.17}$$

solves (4.3) with $u(1, t) \geqq \frac{1}{2} d_1$.

The asserted interior regularity follows from the properties of G and the continuity of u in $\bar{D}_T$. We omit the (standard) arguments.

A similar statement and argument holds for (B_1).

This result allows us to establish a precise version of the statement "If $|u| \leqq M$ on $\bar{D}_T$ and $u(1, T) > 0$ and u is a classical solution of (A_1) on $D_T \cup \Gamma_T$, then u may be continued as a classical solution on $D_{T+\delta} \cup \Gamma_{T+\delta}$ for some $\delta > 0$, with $u(1, t) > 0$ on $[T, T + \delta)$."

Appendix I. We have used, in Theorems 3.2A, B, 3.6A, B, the maximum principle applied to u_t in a rather cavalier fashion. Here we will state and sketch the proof of the precise result. It is similar to that of Lemma 2.3 of [5]. We shall only state it for solutions of (A_1) although it is true for solutions of (B_1) also. A weak form of the first statement was used in Theorems 3.2A, B while second was applied in Theorems 3.6A, B.

THEOREM. *Let f be C^2 on R^1 and g be C^1 except possibly at $u = 0$. Let u solve (A_1) in D_T, be C^2 in x, C^1 in t in D_T and continuous in $D_T \cup \Gamma_T$.*

(a) *If*

(H_1) $\quad u(x, t) > u(x, 0) > 0$ $(0 < u(x, t) < u(x, 0))$ *in $D_T \cup \Gamma_T$ except at $x = 0$,*

then $u_t(x, t) > 0$ (< 0) in D_T and on $x = 1$, $0 < t < T$.

(H_2) (i) *If either* $g'(0)$ *exists or* $u>0$ *on* $x=1$, $0\leqq t<T$,
(ii) u_t *is continuous in* D_T *up to* $t=0$ *on* $(0,1)$,
(iii) $u_x(1,0)=g(u(1,0))$ (*corner compatibility*),
(iv) u_t *is continuous in* D_T *up to* $x=0$ *and* $x=1$ *for* $0<t<T$,
(v) $u_t(x,0)>0$ (<0) *on* $(0,1]$, $u_t(0,0)\geqq 0$ $(\leqq 0)$,

then $u_t(x,t)>0$ (<0) *in* D_T *and on* $x=1$.

Proof. We dispense with the second statement first. Corner compatibility assures us that u_t is continuous in $\bar{D}_{T-\delta}$ for all small $\delta>0$, except possibly at $(0,0)$. Let $v=u_t$. Then we have

$$v_t=v_{xx}+f'(u)v_x+u_xf''(u)v \quad \text{in } D_T,$$
$$v_x=g'(u)v, \qquad x=1, \quad 0<t<T,$$
$$v(0,t)=0, \qquad 0<t<T,$$
$$v(0,x)>0, \qquad 0<x<1,$$

while $v(0,0)\geqq 0$. It follows from arguments similar to those used to prove Lemmas 3.1A, 3.2B that v cannot have a nonpositive minimum in $\bar{D}_{T-\delta}$ except at $(0,0)$. Therefore $v(x,t)>0$ in D_T and on $x=1$.

The proof of the first part is more difficult. To prove it we work in $\bar{D}_{T-\delta}$ and let $0<h<\delta/2$. We let

$$v(x,t)=u(x,t+h)-u(x,t).$$

Then $v(0,t)=0$ if $0<t<T$, $v(x,0)>0$ on $(0,1]$ and v satisfies

$$v_t=v_{xx}+f'(u(x,t+h))v_x+f''(\xi(x,t,h))u_xv$$

where ξ is between $u(x,t+h)$ and $u(x,t)$, while for $0<t\leqq T-\delta$,

$$v_x=g'(\eta(1,t))v$$

where η is between $u(1,t+h)$ and $u(1,t)$. The hypotheses are such that the coefficients of v, v_x are bounded in $\bar{D}_{T-\delta}$ and therefore, by the first and second maximum principles, $v\geqq 0$ in $\bar{D}_{T-\delta}$. It follows that $u_t\geqq 0$ wherever it exists. By interior regularity, u_t exists in D_T and by boundary regularity arguments, u_t exists on $x=1$, $0<t<T$.

Now $\psi=u_t$ satisfies, $\psi\geqq 0$ and

$$\psi_t=\psi_{xx}+f'(u)\psi_x+f''(u)\psi,$$

in $D_{T-\delta}$,

$$\psi_x=g'(u)\psi$$

at $x=1$, $0<t\leqq T-\delta$,

$$\psi(0,t)=0$$

for $0<t\leqq T-\delta$ and

$$\psi(x,0)\geqq 0$$

on $(0,1)$. Therefore $\psi>0$ in $D_{T-\delta}$ unless $\psi\equiv 0$ by the strong maximum principle. But then $\psi\equiv 0$ implies $u(x,0)\equiv u(x,t)$. □

Appendix II. Here we shall establish the signs of $w_1'(\varepsilon)$, $w_0'(\varepsilon)$ along the various stationary solution branches for problems (A), (B). These are crucial in determining the stability of these branches.

In the case of Figs. 1.1 and 1.2, we write (2.7) in the form

$$w = F(w, \varepsilon) = \int_0^1 \frac{d\sigma}{aw^{p-2} + \frac{1}{2}\varepsilon(1-\sigma^2)} \tag{1.1A}$$

where we have agreed to drop the subscript on w_1. When $1 \leqq p < \infty$, $w = w(\varepsilon)$ and we may differentiate implicitly to obtain

$$\frac{dw}{d\varepsilon} = \frac{F\varepsilon}{1 - F_w} \tag{1.2A}$$

we see that

$$F_\varepsilon(w, \varepsilon) = -\frac{1}{2}\int_0^1 \frac{(1-\sigma^2)\, d\sigma}{D^2(\sigma)} < 0 \tag{1.3A}$$

where $D(\sigma)$ is the denominator in the integrand in (1.1A). Using (1.1A) we may calculate F_w in a similar manner. We find that with

$$X = \frac{\varepsilon}{2a} w^{2-p}, \tag{1.4A}$$

we have

$$\frac{dw}{d\varepsilon} = -\frac{w^2 \int_0^1 (1-\sigma^2)\, d\sigma / D^2(\sigma, X)}{2\int_0^1 d\sigma / D(\sigma, X)} \cdot \left[\int_0^1 \frac{[(p-1) + X(1-\sigma^2)]}{D^2(\sigma, X)}\, d\sigma\right]^{-1} \tag{1.5A}$$

provided the integral in brackets does not vanish. We have set

$$D(\sigma, X) = 1 + X(1-\sigma^2). \tag{1.6A}$$

Thus, when $1 \leqq p < \infty$, $w'(\varepsilon) < 0$. Moreover, in this case,

$$\lim_{\varepsilon \to \infty} w(\varepsilon) = 0. \tag{1.7A}$$

This follows from the Dominated Convergence Theorem. Let $\bar{w}$ denote this limit. From (1.1A) we have

$$\bar{w} = \lim_{\varepsilon \to \infty} \int_0^1 \frac{d\sigma}{aw^{p-2}(\varepsilon) + \frac{1}{2}\varepsilon(1-\sigma^2)} = 0$$

if $\bar{w}^{(p-2)} > 0$, which is impossible.

If $0 < p < 1$, the bracketed integral in (1.5A) can vanish. Therefore we view $\varepsilon = \varepsilon(w)$. We know in this case, from the discussion in the example of § 2, that this function is bounded with bound $\varepsilon(p)$ and defined on some interval $[a^{-1/(p-1)}, B)$. However, we may take $B = +\infty$, since as $\varepsilon \to 0$, one of the solution branches $w_1(\varepsilon)$ of equation (2.8) satisfies

$$\lim_{\varepsilon \to 0^+} \varepsilon w_1^{(p-2)} = 0.$$

This tells us that $w_1(\varepsilon) \to +\infty$ as $\varepsilon \to 0$. Thus $B = \infty$ and the variable X in (1.4A) ranges over $(0, \infty)$.

Therefore, we may write

$$\frac{d\varepsilon}{dw} = -\frac{2\int_0^1 (1-\sigma^2)\, d\sigma / D^2(\sigma, X)}{w^2 \int_0^1 d\sigma / D(\sigma, X)} \cdot \left[\int_0^1 \frac{(p-1) + X(1-\sigma^2)}{D^2(\sigma, X)}\, d\sigma\right]. \tag{1.8A}$$

The bracketed integral can (and does) change sign exactly once. Therefore we find that for small X, on the upper branch in Figs. 1.1 and 1.2, $dw/d\varepsilon < 0$ while $dw/d\varepsilon > 0$ on the lower branch.

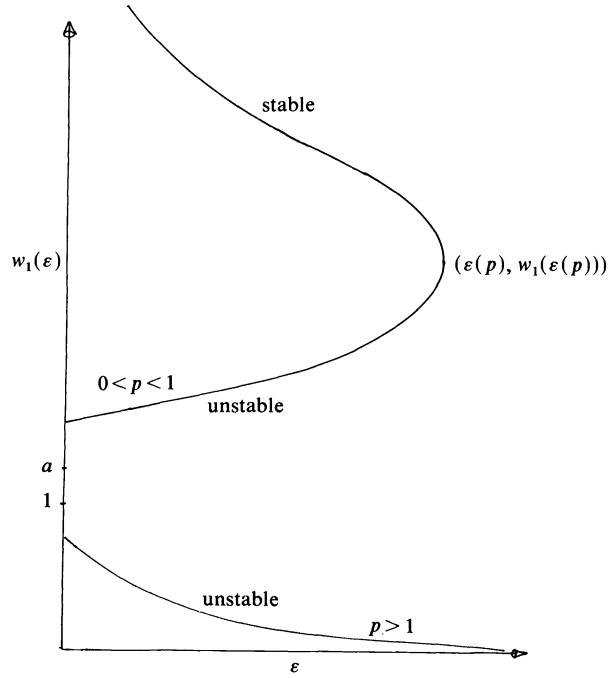


FIG. 1.1. $w_1(\varepsilon)$ *where* $w'(1) = aw^p(1)$, $a > 1$ *and* $w_1(0) = a^{-1/(p-1)}$.

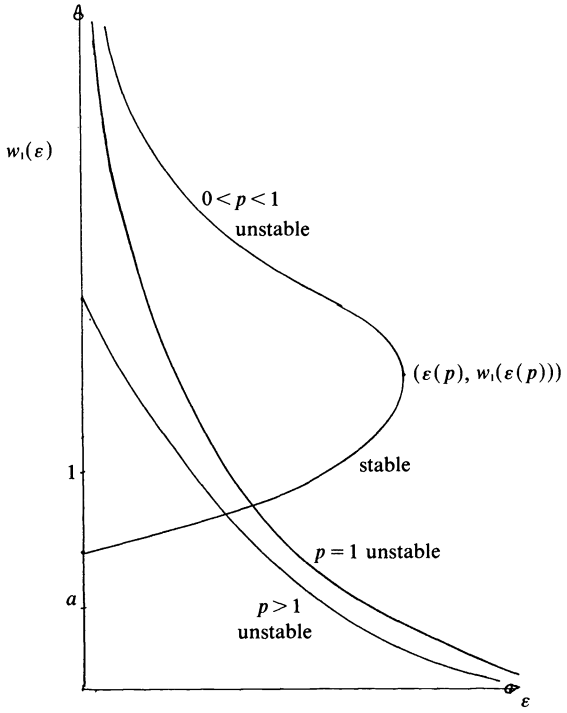


FIG. 1.2. *Same as for Fig.* 1.1 *except* $0 < a < 1$.

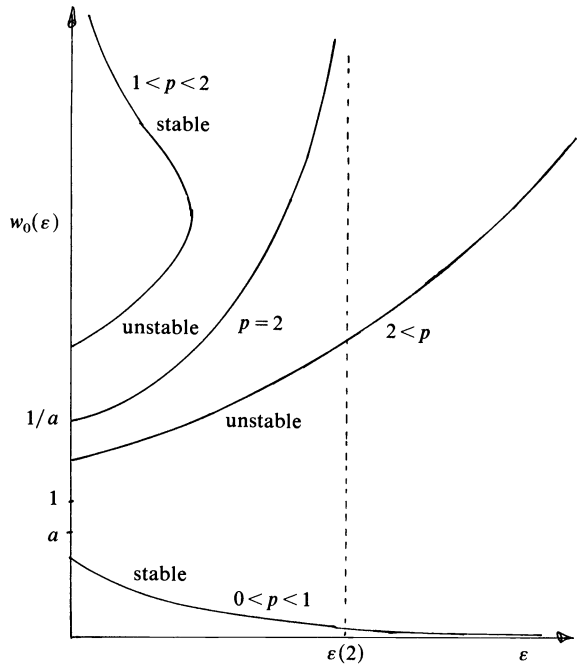


FIG. 2.1. $w_0(\varepsilon)$ *where* $w'(0) = -aw^p(0)$, $0 < a < 1$ *and* $w_0(0) = a^{-1/(p-1)}$.

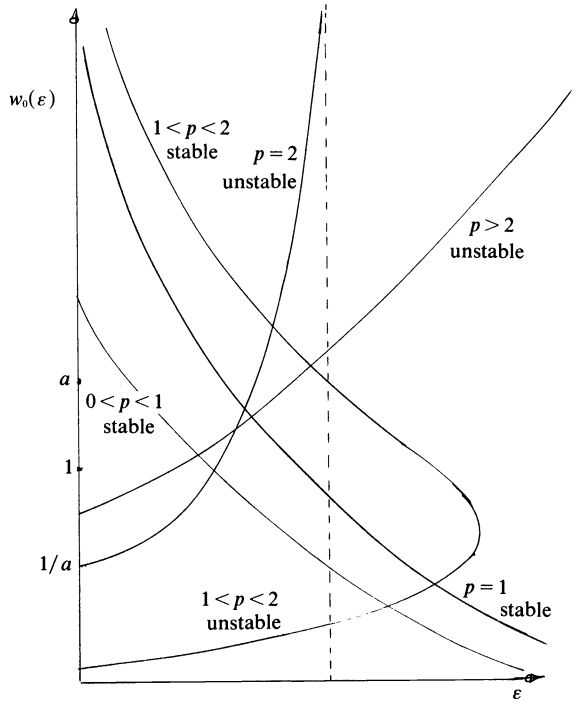


FIG. 2.2. *Same as Fig.* 2.1 *except* $a > 1$.

For problem (B), the argument is similar. In the cases $2 \leqq p < \infty$ and $0 < p \leqq 1$, we have

$$\frac{dw}{d\varepsilon} = -\frac{w^2 \int_0^1 (1-\sigma^2)\, d\sigma / D(\sigma, X)}{2 \int_0^1 d\sigma / D^2(\sigma, X)} \cdot \left[\int_0^1 \frac{(1-p)+X(1-\sigma^2)}{D^2(X, \sigma)}\, d\sigma \right]^{-1}$$

where now $X \in (0, 1)$. Clearly,

$$1-p < (1-p)+X(1-\sigma^2) < 2-p-\sigma^2$$

for X in this range. Thus, if $p \geqq 2$, $dw/d\varepsilon > 0$ while if $0 < p \leqq 1$, $dw/d\varepsilon < 0$. In the case $1 < p < 2$, the argument is similar to that of problem (A) for the case $0 < p < 1$.

It is possible to establish, at least in some cases, the precise curvatures in Figs. 1.1, 1.2, 2.1 and 2.2. However, such an analysis appears to add nothing to the stability results so we omit it.

Acknowledgments. The author would like to thank A. Majda for posing the question of blow up for solutions of (C). He also thanks H. Amann, J. Anderson, G. Lieberman and P. Sacks for several useful conversations, and is indebted to the referee for bringing [13], [14] to his attention.

REFERENCES

[1] T. F. CHEN, H. A. LEVINE AND P. E. SACKS, *Analysis of a convective reaction-diffusion equation* (I), to appear.

[2] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[3] M. W. HIRSCH, *Differential equations and convergence almost everywhere of strongly monotone semiflows*, Contemp. Math., 17 (1983), pp. 267-285.

[4] V. K. KALANTAROV, *Collapse of the solutions of parabolic and hyperbolic equations with nonlinear boundary conditions*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov, 127 (1983), pp. 75-83. (In Russian.) J. Soviet Math., 27 (1984), pp. 2601-2606. (In English.)

[5] H. A. LEVINE, *The quenching of solutions of linear parabolic and hyperbolic equations with nonlinear boundary conditions*, this Journal, 14 (1983), pp. 1139-1153.

[6] H. A. LEVINE AND R. A. SMITH, *A potential well theory for the heat equation with a nonlinear boundary condition*, Math. Methods Appl. Sci., to appear.

[7] ———, *A potential well theory for the wave equation with a nonlinear boundary condition*, J. Reine Angew. Math., 374 (1986), pp. 1-23.

[8] H. A. LEVINE AND L. E. PAYNE, *Nonexistence theorems for the heat equation with nonlinear boundary conditions and for the porous medium equation backward in time*, J. Differential Equations, 16 (1974), pp. 319-334.

[9] H. MATANO, *Asymptotic behavior and stability of solutions of semilinear diffusion equations*, Publ. Res. Inst. Math. Sci., 15 (1979), pp. 401-454.

[10] ———, *Existence of nontrivial unstable sets for equilibriums of strongly order preserving systems*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math., 30 (1984), pp. 645-673.

[11] L. E. PAYNE AND D. H. SATTINGER, *Saddle points and instability of nonlinear hyperbolic equations*, Arch. Rational Mech. Anal., 30 (1968), pp. 148-172.

[12] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[13] W. WALTER, *On existence and nonexistence in the large solutions of parabolic differential equations with a nonlinear boundary condition*, this Journal, 6 (1975), pp. 85-90.

[14] H. AMANN, *Quasilinear parabolic systems under nonlinear boundary conditions*, Arch. Rational Mech. Anal., 92 (1986), pp. 153-192.

[15] H. A. LEVINE, L. E. PAYNE, P. E. SACKS AND B. STRAUGHAN, *Analysis of a Convective Reaction Diffusion Equation* (II), to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

006

STRONG SOLUTIONS OF A QUASILINEAR WAVE EQUATION WITH NONLINEAR DAMPING*

DANG DINH ANG† AND A. PHAM NGOC DINH‡

Abstract. We study the following initial and boundary value problem

$$u_{tt} - \Delta u_t - \sum_{i=1}^{N} \sigma_i'(u_{x_i}) u_{x_i x_i} + |u_t|^\alpha \operatorname{sgn} u_t = 0,$$

$$0 < \alpha < 1 \quad (x, t) \quad \text{in } \Omega \times]0, T[,$$

$$u = 0 \quad \text{on } \partial\Omega,$$

$$u(x, 0) = u_0(x), \qquad u_t(x, 0) = u_1(x)$$

where Ω is a bounded domain in R^N with a sufficiently regular boundary $\partial\Omega$. In § 1, it is proved that for u_0 in $H_0^1(\Omega)$, u_1 in $L_2(\Omega)$, σ_i in $C(\mathbb{R}, \mathbb{R})$ nondecreasing and inducing mappings of $L_2(\Omega)$ into itself, taking bounded sets into bounded sets, the problem admits a global weak solution. If, in addition, the σ_i's are assumed locally Lipschitzian, then the solution is unique.

In § 2, it is proved that for $N = 1$, u_0 in $H_0^1(\Omega) \cap H^2(\Omega)$, u_1 in $L_2(\Omega)$ and σ_i in $C^1(\mathbb{R}, \mathbb{R})$ with $\sigma_i' > 0$ nondecreasing and locally Hölder continuous, there exists a unique strong solution u of the initial and boundary value problem with the following properties: $t \to u(t)$ is continuous on $t \geqq 0$ to $H_0^1(\Omega) \cap H^2(\Omega)$, is continuously differentiable on $t > 0$ to $H_0^1(\Omega) \cap H^2(\Omega)$, is continuously differentiable on $t \geqq 0$ to $L_2(\Omega)$, and is twice continuously differentiable on $t > 0$ to $L_2(\Omega)$.

Key words. strong solutions, nonlinear PDE, analytic semi-groups

AMS(MOS) subject classifications. 35, 35B, 35K, 35L, 47H

Introduction. We shall consider the following nonlinear initial and boundary value problem

$$u_{tt} - \Delta u_t - \frac{\partial}{\partial x_i} \sigma_i(u_{x_i}) + f(u_t) = 0, \qquad (x, t) \in \Omega \times]0, T[, \tag{0.1}$$

$$u = 0 \quad \text{on } \partial\Omega, \tag{0.2}$$

$$u(x, 0) = u_0(x), \qquad u_t(x, 0) = u_1(x) \tag{0.3}$$

where

$$f(u_t) = |u_t|^\alpha \operatorname{sgn} u_t, \qquad 0 < \alpha < 1, \qquad u_t = \frac{\partial u}{\partial t}, \qquad u_{x_i} = \frac{\partial u}{\partial x_i}.$$

In (0.1), Ω is a bounded domain in R^N with a sufficiently smooth boundary $\partial\Omega$, and σ_i, $i = 1, \cdots, N$ are continuous functions, satisfying certain monotonic and other conditions to be specified later.

Equations of the type (0.1), with $f = 0$, were given the first systematic treatment by Greenberg, MacCamy and Mizel [9] in the case of space dimension $N = 1$. They were proposed by the authors [6] as field equations governing the longitudinal motion of a viscoelastic bar obeying the nonlinear Voight model. Since the appearance of the work of Greenberg, MacCamy and Mizel, there has been a rather impressive literature on equations of the type (0.1) above, e.g., Caughey and Ellison [2], Clements [3], [4], Dafermos [5], Kozhanov, Larkin and Janenko [10], Yamada [16], [17], to name but a few. Of particular relevance to the present paper are the works of Clements and of

* Received by the editors December 21, 1984; accepted for publication (in revised form) March 1, 1987.
† Department of Mathematics, Dai Hoc Tong Hop, Ho Chi Minh City University, Viet Nam.
‡ Département de Mathématiques, Université d'Orléans, 45046 Orléans Cedex, France.

Webb [15]. The paper of Yamada [17] should also be consulted for reference. The remainder of the paper consists of two sections. In § 1, it is proved that for each given interval $]0, T[$, there exists a unique weak solution of (0.1)-(0.3) on the interval, under certain monotonic and other conditions on the σ_i's.

The method used here is a combination of compactness and monotonicity; such a method has enabled us to avoid the smoothness conditions as imposed, e.g., in Clements [3] (while the solution obtained here is weaker). For uniqueness, the σ_i's should satisfy certain contractive properties. In § 2, it is proved that for $N=1$, u_0 in $H_0^1(\Omega)\cap H^2(\Omega)$, u_1 in $L_2(\Omega)$ and σ_i in $C^1(\mathbb{R},\mathbb{R})$ with σ_i' locally Hölder continuous and positive, a unique strong solution of (0.1)-(0.3) exists for $t>0$. Note that the derivatives σ_i' are not used to satisfy any Lipschitzian condition. As a consequence, the usual method using semigroup theory to formulate the problem (0.1)-(0.3) as an integral equation and solve it by means of the contraction principle would not work in our case. Instead, we have found it efficient to use a combination of Galerkin approximation and analytic semigroup theory. The Galerkin method gives us a weak solution, which will be proved to be a strong solution, using Pazy's results on analytic semigroups [13].

1. Let

$$L_2 = L_2(\Omega), \quad H_0^1 = H_0^1(\Omega), \quad H^2 = H^2(\Omega).$$

Here $H_0^1(\Omega)$ and $H^2(\Omega)$ denote the usual Sobolev spaces on Ω. Let $\langle\ \cdot\ \rangle$ denote either the L_2-inner product or the paring of a continuous linear functional with an element of a function space. Let $\|\cdot\|_X$ be a norm on a Banach space X and let X^* be its dual. We denote by $L_p(0, T; X)$, $1\leqq p\leqq\infty$, the space of functions f on $[0, T]$ to X such that

$$\|f\|_{L_P(0,T;X)} = \left(\int_0^T \|f(t)\|_X^P\, dt\right)^{1/p} \quad \text{for } 1\leqq p<\infty,$$

$$\|f\|_{L_\infty(0,T;X)} = \operatorname{ess\,sup} \|f(t)\|_X \quad \text{for } p=\infty.$$

Then we have the following.

THEOREM 1 (Weak solution). *Let σ_i, $i=1,\cdots,N$ be real-valued functions satisfying the following*:

(1.1) *σ_i in $C(\mathbb{R},\mathbb{R})$, nondecreasing, $\sigma_i(0)=0$ each $\bar\sigma_i: L_2\to L_2$, where $\bar\sigma_i f=\sigma_i\circ f$ for f in L_2*

(1.2) *takes bounded sets into bounded sets and*

(1.3) *is locally Lipschitzian*

(1.4) *u_0 in H_0^1 and u_1 in L_2.*

Then for each $T>0$, the initial and boundary value problem (0.1)-(0.3) *admits a unique weak solution $u(\cdot)$ on $]0, T[$ with the following properties*:

(1.5) *u in $L_\infty(0, T; H_0^1)$ and u_t in $L_\infty(0, T; L_2)\cap L_2(0, T; H_0^1)$,*

(1.6) *$u(\cdot)$ locally Hölder continuous on $[0, T[$ to H_0^1.*

Remark 1. If σ_i is in $C^1(R, R)$ with a bounded derivative as in [3], then obviously σ_i satisfies (1.2).

Proof. The proof is a combination of compactness and monotony arguments, and consists of several steps. Step 1 is devoted to constructing Galerkin approximations and establishing a priori estimates. Step 2 is concerned with existence of the solution,

which will result from appropriate limiting processes. The final Step 3 will settle the question of uniqueness.

Step 1. Consider a special basis of H_0^1: $w_1, \cdots, w_n, \cdots$ formed by the eigenfunctions of Laplacian Δ on H_0^1. Let $(w_1, \cdots, w_n)$ be the linear space generated by $w_1, \cdots, w_n$. Let

$$u^{(n)}(t) = \sum_{k=1}^{n} c_{k,n}(t) w_k \tag{1.7}$$

be a solution of the following system:

$$\begin{aligned} &\langle u_{tt}^{(n)}(t), w_k\rangle + \sum_{i=1}^{N} \langle \sigma_i(u_{x_i}^{(n)}(t)), w_{k,x_i}\rangle - \langle \Delta u_t^{(n)}(t), w_k\rangle \\ &\qquad + \langle f(u_t^{(n)}(t)), w_k\rangle = 0, \end{aligned} \tag{1.8}$$

$$u^{(n)}(0) = u_{0n}, \qquad u_t^{(n)}(0) = u_{1n} \tag{1.9}$$

where $1 \leqq k \leqq n$

$$u_{0n} \to u_0 \quad \text{in } H_0^1,$$

$$u_{1n} \to u_1 \quad \text{in } L_2.$$

Note that such a solution $u^{(n)}(t)$ clearly exists on a sufficiently small interval $[0, T_n[$. Note also that the a priori estimates which follow allow us to take T_n equal to T.

A priori estimates. Let us multiply each equation of (1.8) by $\dot{c}_{k,n}(t)$, sum up and integrate with respect to the time variable from 0 to t. Then, we shall have, after some rearrangements

$$\begin{aligned} &\|u_t^{(n)}(t)\|^2 + 2\int_0^t \|\nabla u_t^{(n)}(s)\|^2\, ds + 2\sum_1^N \int_\Omega h_i(u_{x_i}^{(n)}(t))(x)\, dx \\ &\qquad + 2\int_0^t \langle f(u_t^{(n)}(s)), u_t^{(n)}(s)\rangle\, ds \\ &\qquad = \|u_{1n}\|^2 + 2\sum_1^N \int_\Omega h_i(u_{0n,x_i})(x)\, dx \end{aligned} \tag{1.10}$$

where

$$h_i(z) = \int_0^z \sigma_i(s)\, ds, \tag{1.11}$$

since we have the formula

$$\frac{d}{dt}\int_\Omega h_i(u_{x_i}^{(n)}(t))\, dx = -\left\langle \frac{\partial}{\partial x_i}\sigma_i(u_{x_i}^{(n)}(t)), u_t^{(n)}(t)\right\rangle,$$

$u_t^{(n)}(x, t)$ being equal to 0 on $\partial\Omega$.

By (1.1)

$$\int_\Omega h_i(u_{x_i}^{(n)}(t))(x)\, dx \geqq 0, \qquad i = 1, \cdots, N. \tag{1.12}$$

By (1.2),

$$0 \leqq \int_\Omega h_i(u_{0n,x_i})(x)\, dx \leqq C \tag{1.13}$$

where C is a constant independent of n.

Noting that

$$\langle f(v), v\rangle \geqq 0 \tag{1.14}$$

it follows from (1.10) that

$$\|u_t^{(n)}(t)\|^2 + \int_0^t \|\nabla u_t^{(n)}(s)\|^2\, ds \leqq C, \qquad 0 \leqq t \leqq T_n \tag{1.15}$$

where C is independent of n.

From (1.15), we deduce that, in particular, $t \to \nabla u^{(n)}(t)$ is Hölder continuous on $[0, T]$ and that

$$\|\nabla u^{(n)}(t)\| \leqq M_T, \qquad 0 \leqq t \leqq T \quad \text{for all } n \tag{1.16}$$

where M_T depends on T.

Define

$$\begin{aligned} &A: H_0^1 \to H^{-1}(\Omega) = H^{-1} \quad \text{(the dual of } H_0^1\text{)},\\ &Au: -\sum_1^N \sigma_i(u_{x_i})_{x_i}. \end{aligned} \tag{1.17}$$

Then

$$\langle Au, v\rangle = \sum_1^N \langle \sigma_i(u_{x_i}), v_{x_i}\rangle, \qquad u, v \text{ in } H_0^1.$$

Therefore,

$$|\langle Au^{(n)}(t), v\rangle| \leqq \left(\sum_1^N \|\sigma_i(u_{x_i}^{(n)}(t))\|^2\right)^{1/2} \|v\|_{H_0^1}. \tag{1.18}$$

By (1.16) and (1.2), it follows that

$$\|Au^{(n)}(t)\|_* \leqq M_T, \qquad 0 \leqq t \leqq T \tag{1.19}$$

where $\|\cdot\|_*$ is the dual norm on H^{-1}.

On the other hand, we have

$$\|f(u_t^{(n)}(t)\| \leqq C, \qquad 0 \leqq t \leqq T \tag{1.20}$$

(since Ω is bounded and $0 < \alpha < 1$).

We need an estimate on the $u_{tt}^{(n)}$. From the approximated problem

$$u_{tt}^{(n)}(t) - \Delta u_t^{(n)}(t) + Au^{(n)}(t) + f(u_t^{(n)}(t)) = 0, \tag{1.21}$$

we can deduce the following for each $v \in H_0^1$:

$$|\langle u_{tt}^{(n)}(t), v\rangle| \leqq (\|\nabla u_t^{(n)}(t)\| + \|Au^{(n)}(t)\|_* + \|f(u_t^{(n)}(t))\|)\|v\|_{H^1}. \tag{1.22}$$

By (1.15), (1.19) and (1.20) it follows that

$$\int_0^T \|u_{tt}^{(n)}(t)\|_{H^{-1}}^2\, dt \leqq 2\int_0^T (\|\nabla u_t^{(n)}(t)\|^2 + M_T)\, dt \leqq M_T. \tag{1.23}$$

The M_T always indicates constants independent of n, but depending on T. Therefore,

$$\int_0^T \|u_{tt}^{(n)}(t)\|_{H^{-1}}^2\, dt \leqq M_T, \qquad 0 \leqq t \leqq T \quad \text{for all } n. \tag{1.24}$$

Consequently,

$$\int_0^T \|\Delta u_t^{(n)}(t)\|_{H^{-1}}^2 \, dt \leqq M_T, \qquad 0 \leqq t \leqq T \quad \text{for all } n. \tag{1.25}$$

By (1.16), (1.15), (1.24), (1.19) and (1.20), we can extract a subsequence of $u^{(n)}$ still denoted by $u^{(n)}$, such that

$$u^{(n)} \to u \text{ in } L_\infty(0, T; H_0^1) \text{ weak } *, \tag{1.26}$$

$$\nabla u_t^{(n)} \to \nabla u_t \text{ in } L_2(0, T; L_2) \text{ weak}, \tag{1.27}$$

$$u_{tt}^{(n)} \to u_{tt} \text{ in } L_2(0, T; H^{-1}) \text{ weak}, \tag{1.28}$$

$$Au^{(n)} \to \xi \text{ in } L_\infty(0, T; H^{-1}) \text{ weak } *, \tag{1.29}$$

$$f(u_t^{(n)}) \to \chi \text{ in } L_\infty(0, T; L_2) \text{ weak } *. \tag{1.30}$$

Using a lemma on compactness [12] applied to (1.26) and (1.27), on the one hand, and to (1.27) and (1.28), on the other hand, we can extract from the sequence $\{u^{(n)}\}$ a subsequence, still denoted by $\{u^{(n)}\}$, such that

$$u^{(n)} \to u \text{ strongly in } L_2(Q) \text{ with } Q = \Omega \times]0, T[, \tag{1.31}$$

$$u_t^{(n)} \to u_t \text{ strongly in } L_2(Q). \tag{1.32}$$

Step 2. *Existence of a solution through a limiting process.* Letting $n \to \infty$ in (1.8), we find from (1.26)-(1.30) that u satisfies the equation

$$\begin{aligned} &\frac{d}{dt}\langle u_t(t), v\rangle + \langle \nabla u_t(t), \nabla v\rangle + \langle \xi(t), v\rangle + \langle \chi(t), v\rangle \\ &\quad = 0 \quad \text{a.e. } t \text{ in }]0, T[\quad \text{for all } v \text{ in } H_0^1. \end{aligned} \tag{1.33}$$

The initial conditions are satisfied since

(1.34) $u^{(n)}$ and u are in $C(0, T; L_2)$ implying that $u_n(0) = u_{0n} \to u(0)$ strongly in L_2 and hence $u(0) = u_0$, and

(1.35) $\langle u_t^{(n)}(t), w_k\rangle$ and $\langle u_t(t), w_k\rangle$ are in $C(0, T; \mathbb{R})$ implying that $\langle u_t^{(n)}(0) - u_t(0), w_k\rangle \to 0$ for $n \to \infty$ and hence $u_t(0) = u_1$.

So, we shall have proved the existence of a weak or distributional solution of (0.1)-(0.3) once we have shown that

$$\xi = Au \quad \text{and} \quad \chi = f(u_t). \tag{1.36}$$

We shall require the following two lemmas the proofs of which are immediate.

LEMMA 1. *The operator A defined in* (1.17) *is a monotone and hemicontinuous operator from* H_0^1 *to* H^{-1}, *i.e.*,

(i) $\langle Au - Av, u - v\rangle \geqq 0$ *for all* u, v *in* H_0^1,

(ii) *the map* $s \to \langle A(u + sv), w\rangle$ *is continuous on* $\mathbb{R}$ *to* $\mathbb{R}$.

LEMMA 2. *The function* $u \to f(u)$ *generates a monotone and hemicontinous operator on* L_2 *to* L_2.

We can now complete the existence proof.

Let

$$\xi_n(t) = \int_0^t \langle Au^{(n)}(s) - Av(s), u^{(n)}(s) - v(s)\rangle \, ds, \qquad 0 \leqq t < T. \tag{1.37}$$

We have, since A is monotone,

$$\xi_n \geqq 0 \quad \text{for all } v \text{ in } L_2(0, T; H_0^1).$$

From the relation

$$\begin{aligned}\int_0^t \langle Au^{(n)}(s), u^{(n)}(s)\rangle\, ds = &-\int_0^t \langle u_{tt}^{(n)}(s), u^{(n)}(s)\rangle\, ds \\ &-\frac{1}{2}\|\nabla u^{(n)}(t)\|^2+\frac{1}{2}\|u_{0n}\|^2_{H_0^1} \\ &-\int_0^t \langle f(u_t^{(n)}(s)), u^{(n)}(s)\rangle\, ds\end{aligned} \tag{1.38}$$

we obtain by passing to the limit

$$\begin{aligned}&\limsup_{n\to\infty} \int_0^t \langle Au^{(n)}(s), u^{(n)}(s)\rangle\, ds \\ &\quad\leqq -\frac{1}{2}\|\nabla u(t)\|^2+\frac{1}{2}\|u_0\|^2_{H_0^1} \\ &\quad\quad -\int_0^t \langle \chi(s), u(s)\rangle\, ds+\limsup_{n\to\infty}\left\{-\int_0^t \langle u_{tt}^{(n)}(s), u^{(n)}(s)\rangle\right\} ds\end{aligned} \tag{1.39}$$

a.e. t in $]0, T[$.

u satisfies the equation

$$u_{tt}-\Delta u_t+\xi+\chi=0, \tag{1.40}$$

which has a meaning in $L_2(0, T; H^{-1})$ as can be seen from (1.25) and (1.28)-(1.30).

Taking the inner product of (1.40) with u which belongs to $L_\infty(0, T; H_0^1)$, integrating with respect to the time variable from s to t, we obtain

$$\begin{aligned}&\int_s^t \langle u_{tt}(t), u(t)\rangle\, dt+\int_s^t \langle \xi(t), u(t)\rangle\, dt \\ &\quad=\frac{1}{2}(\|\nabla u(s)\|^2-\|\nabla u(t)\|^2)-\int_s^t \langle \chi(t), u(t)\rangle\, dt.\end{aligned} \tag{1.41}$$

From (1.41) we have, passing to the limit as $s\to 0$:

$$\begin{aligned}&\int_0^t \langle u_{tt}(t), u(t)\rangle\, dt+\int_0^t \langle \xi(t), u(t)\rangle\, dt \\ &\quad\geqq\frac{1}{2}(\|u_0\|^2_{H_0^1}-\|\nabla u(t)\|^2)-\int_0^t \langle \chi(t), u(t)\rangle\, dt.\end{aligned} \tag{1.42}$$

Then, by virtue of (1.42), the inequality (1.39) can be rewritten as

$$\begin{aligned}\limsup_{n\to\infty} \int_0^t \langle Au^{(n)}(s), u^{(n)}(s)\rangle\, ds \leqq &\int_0^t \langle \xi(s), u(s)\rangle\, ds \\ &+\int_0^t \langle u_{tt}(s), u(s)\rangle\, ds \\ &+\limsup_{n\to\infty}\left(-\int_0^t \langle u_{tt}^{(n)}(s), u^{(n)}(s)\rangle\, ds\right)\end{aligned} \tag{1.43}$$

a.e. t in $]0, T[$.

Integrating by parts, we have the relation

$$\int_0^t \langle u_{tt}^{(n)}(s), u^{(n)}(s)\rangle\, ds = -\int_0^t \|u_t^{(n)}(s)\|^2\, ds + \langle u^{(n)}(t), u_t^{(n)}(t)\rangle - \langle u_{0n}, u_{1n}\rangle \quad \text{for all } t \text{ in } [0, T[. \tag{1.44}$$

Passing to the limit in (1.44), we finally obtain

$$\lim_{n\to\infty} \int_0^t \langle u_{tt}^{(n)}(s), u^{(n)}(s)\rangle\, ds = -\int_0^t \|u_t(s)\|^2\, ds + \langle u(t), u_t(t)\rangle - \langle u_0, u_1\rangle = \int_0^t \langle u_{tt}(s), u(s)\rangle\, ds \quad \text{a.e. } t \text{ in } [0, T[\tag{1.45}$$

since

$$\lim_{n\to\infty} \langle u^{(n)}(t), u_t^{(n)}(t)\rangle = \langle u(t), u_t(t)\rangle \quad \text{a.e. } t \text{ in } [0, T[\tag{1.46}$$

as can be seen from (1.31) and (1.32).

Therefore,

$$\limsup_{n\to\infty} \int_0^t \langle Au^{(n)}(s), u^{(n)}(s)\rangle\, ds \leqq \int_0^t \langle \xi(s), u(s)\rangle\, ds \quad \text{a.e. } t \text{ in } [0, T[\tag{1.47}$$

which implies that

$$0 \leqq \int_0^T \langle \xi(t) - Av(t), u(t) - v(t)\rangle\, dt \quad \text{for all } v \text{ in } L_2(0, T; H_0^1). \tag{1.48}$$

In (1.48), if we choose $v = u - sw$ with $s \geqq 0$ and w in $L_2(0, T; H_0^1)$ and use the hemicontinuity of A, then we get

$$\xi = Au. \tag{1.49}$$

Now, from (1.30) and (1.32) we have

$$\lim_{n\to\infty} \int_0^T \langle f(u_t^{(n)}(s)), u_t^{(n)}(s)\rangle\, ds = \int_0^T \langle \chi(s), u_t(s)\rangle\, ds. \tag{1.50}$$

Using Lemma 2, we conclude that

$$\chi = f(u_t). \tag{1.51}$$

The existence proof is completed. Now, it is clear that

$$\|\nabla u(t) - \nabla u(t')\| \leqq |t - t'|^{1/2} \left(\int_0^T \|\nabla u_t(s)\|^2\, ds \right)^{1/2} \tag{1.52}$$

and hence, that $t \to u(t)$ is Hölder continuous from $[0, T[$ to H_0^1.

Step 3. *Uniqueness proof.* Let u and v be two weak solutions of the problem (0.1)-(0.3). Then $w = u - v$ is a weak solution of the problem

$$w_{tt} - \Delta w_t - \Delta w + f(u_t) - f(v_t) = \sum_{i=1}^N (\sigma_i(u_{x_i}) - \sigma_i(v_{x_i}))_{x_i} - \Delta w, \tag{1.53}$$

$$w(0) = 0 = w_t(0), \tag{1.54}$$

$$w = 0 \quad \text{on } \partial\Omega. \tag{1.55}$$

Multiplying (1.53) by w_t and integrating over Ω, we obtain

$$\frac{1}{2}\frac{d}{dt}(\|w_t(t)\|^2+\|\nabla w(t)\|^2)+\|\nabla w_t(t)\|^2+\langle f(u_t(t))-f(v_t(t)),\, u_t(t)-v_t(t)\rangle$$
$$(1.56)\qquad =\sum_{i=1}^{N}\langle \sigma_i(u_{x_i}(t))-\sigma_i(v_{x_i}(t)),\, w_{tx_i}(t)\rangle+\langle \nabla w(t), \nabla w_t(t)\rangle.$$

Using (1.3) and the monotonicity of f, we get from (1.56)

$$(1.57)\qquad \frac{1}{2}\frac{d}{dt}(\|\nabla w_t(t)\|^2+\|\nabla w(t)\|^2)\leqq a\|\nabla w(t)\|^2$$

for some $a>0$. From (1.57) we deduce that

$$(1.58)\qquad \|w_t(t)\|^2+\|\nabla w(t)\|^2=0$$

Thus $w=0$.

The proof of the theorem is completed.

2. We shall consider the problem of global existence of strong solutions of (0.1)-(0.3). To this end, we shall have to strengthen conditions on the initial data and on the σ_i's. The role of the space dimension is important in this connection, and, in order to simplify matters, we shall limit ourselves to $N=1$. With the latter restriction, it will be sufficient, from our point of view, to place the following requirements on the initial data on the coefficients of the field equation:

$$(2.1)\qquad u_0 \quad \text{in } H_0^1\cap H^2, \qquad u_1 \quad \text{in } L_2,$$

$$(2.2)\qquad \sigma_i \text{ in } C^1(\mathbb{R},\mathbb{R}),\ \sigma_i'>0,\ \sigma_i(0)=0 \text{ and } \sigma_i' \text{ locally Hölder continuous.}$$

Then, we have the following.

THEOREM 2. *Let $N=1$ and let* (2.1) *and* (2.2) *hold. Then, there exists a unique solution $u(\cdot)$ of* (0.1)-(0.3) *with the following properties*:

$$(2.3)\qquad t\to u(t) \text{ is continuously on } t\geqq 0 \text{ to } H_0^1\cap H^2, \text{ continuously differentiable on } t>0 \text{ to } H_0^1\cap H^2;$$

$$(2.4)\qquad t\to u(t) \text{ is continuously differentiable on } t\geqq 0 \text{ to } L_2 \text{ and twice continuously differentiable on } t>0 \text{ to } L_2;$$

$$(2.5)\qquad t\to \Delta u(t) \text{ is continuously differentiable on } t>0 \text{ to } L_2.$$

Proof. The idea of the proof is as follows. We take the weak solution w of (0.1)-(0.3) (which will presently be shown to exist), and then, using the analytic theory of semigroup, we shall prove that w is in fact the strong solution of the theorem. The proof consists of two steps. In Step 1, we shall establish the existence of a unique weak solution on R_+. In Step 2, we shall prove that such a solution is in fact a strong solution.

Step 1. We shall only sketch the proof since it is very similar to (in fact considerably simpler than) the proof of theorem 1. According to the a priori estimates already obtained in the proof of Theorem 1, we had

$$(2.6)\qquad \|u_t^{(n)}(t)\|\leqq M \quad \text{for all } n \text{ and } 0\leqq t\leqq T_n,$$

$$(2.7)\qquad \int_0^t \|\nabla u_t^{(n)}(s)\|^2\, ds\leqq M \quad \text{for all } n \text{ and } 0\leqq t\leqq T_n$$

where $u^{(n)}(t)$ is a Galerkin approximate solution of (0.1)-(0.3) as in the proof of Theorem 1, and $[0, T_n[$ is its interval of existence. Now, using the hypothesis that u_0 be in $H_0^1 \cap H^2$, we have the following bound on $\Delta u^{(n)}$:

$$\|\Delta u^{(n)}(t)\| \leqq M \quad \text{for all } n \text{ and } 0 \leqq t \leqq T_n. \tag{2.8}$$

Note that we can take $T_n = T$ for all n, where T is an arbitrary positive number. Since $N = 1$, we have

$$\|\nabla u^{(n)}(t)\|_\infty \leqq K \|\Delta u^{(n)}(t)\| \quad \text{for all } t \leqq T \tag{2.9}$$

where K is independent of n and t; from (2.9) and (2.8) we deduce that, in particular.

$$\|\nabla u^{(n)}(t)\|_\infty \leqq KM \quad \text{for all } t \text{ in } [0, T]. \tag{2.10}$$

By passing to the limit much as in the proof of Theorem 1, this time using the inequality (2.10), we conclude that there exists a weak solution w of (0.1)-(0.3) on $]0, T[$ with the following properties:

$$\|\Delta w(t)\| \leqq M \quad \text{for all } 0 \leqq t \leqq T, \tag{2.11}$$

$$\|w_t(t)\| \leqq M \quad \text{for all } 0 \leqq t \leqq T, \tag{2.12}$$

$$\int_0^t \|\nabla w_t(s)\|^2 \, ds \leqq M \quad \text{for all } 0 \leqq t \leqq T. \tag{2.13}$$

Uniqueness is proved in a standard manner, much as in the proof of Theorem 1. We conclude that there exists a unique weak solution w on R_+ with the same bounds as in (2.11)-(2.13).

Step 2. Let w be the weak solution of (0.1)-(0.3) on $\mathbb{R}_+$ as in Step 1. Then w is the solution of the following (equivalent) equation

$$w_t = \Delta w + u_1 - \Delta u_0 + G(w) + F(w) \tag{2.14}$$

where

$$G(w(t)) = \int_0^t H(w(s)) \, ds, \qquad F(w(s)) = -\int_0^t f(w_t(s)) \, ds. \tag{2.15}$$

Here

$$H(w) = \sigma'(w_x) w_{xx}. \tag{2.16}$$

Since (2.14) is equivalent to (0.1)-(0.3), w is its unique solution. Consider the differential equation

$$u_t = \Delta u + u_1 - \Delta u_0 + G(w) + F(w) \tag{2.17}$$

with the initial condition

$$u(0) = u_0. \tag{2.18}$$

Let $S(t)$ be the semigroup on L_2 generated by the Laplacian $\Delta(\operatorname{dom} \Delta = H_0^1 \cap H^2)$. Note that $S(t)$ is in fact an analytic semi-group, a fact which will be crucial for what follows. The solution of (2.17)-(2.18) is given by

$$\begin{aligned} u(t) = S(t)u_0 + \int_0^t S(t-s)(u_1 - \Delta u_0) \, ds \\ + \int_0^t S(t-s)(G(w(s)) + F(w(s))) \, ds. \end{aligned} \tag{2.19}$$

From (2.19), we get

$$\Delta u(t) = S(t)\Delta u_0 + S(t)(u_1 - \Delta u_0) - u_1 + \Delta u_0 + \int_0^t S(t-s)(H(w(s)) - F(w_t(s)))\, ds - G(w(t)) - F(w(t)). \tag{2.20}$$

We note from (2.20) that $\Delta u(t)$ is continuous on $t \geqq 0$, and hence, by (2.17) that $u_t(t)$ is continuous on $t \geqq 0$. From now on we shall use freely Pazy's results on analytic semigroups [13]. Since $S(t)$ is an analytic semigroup, we deduce from (2.20) that $t \to \Delta u(t)$ is Hölder continuous, and hence, by (2.17), $t \to u_t(t)$ is also Hölder continuous on $t > 0$ to L_2. Now, by the uniqueness of w as a solution of (2.14), we see that $u = w$. Thus, $w(\cdot)$ and $w_t(\cdot)$ are Hölder continuous. Thus $f(w_t(\cdot))$ is Hölder continuous. By (2.9), the Hölder continuity of σ_i' and by the Hölder continuity of $\Delta w(\cdot)$ noted above, it follows that $H(w(\cdot))$ is Hölder continuous. Thus, it follows from Pazy's results that $t \to \Delta w(t)$ is continuously differentiable on $t > 0$ to L_2, and that, similarly by (2.17), $t \to w_t$ is continuously differentiable on $t > 0$ to L_2. Thus w is the (unique) strong solution of (0.1)-(0.3) in the sense that it satisfies (2.3)-(2.5) (with $w = u$). This completes the proof of the theorem.

Remark 2. It was established in the course of the proof that the strong solution w satisfies the following inequalities (cf. (2.11)-(2.13)):

$$\|\Delta w(t)\| \leqq M \quad \text{for all } t \geqq 0, \tag{2.21}$$

$$\|w_t(t)\| \leqq M \quad \text{for all } t \geqq 0, \tag{2.22}$$

$$\int_0^t \|\nabla w_t(s)\|^2\, ds \leqq M \quad \text{for all } t \geqq 0. \tag{2.23}$$

Acknowledgement. The authors wish to thank the referee of this paper for his constructive and useful remarks.

REFERENCES

[1] F. E. BROWDER, *Nonlinear equations of evolution*, Ann. of Math., 80 (1984), pp. 485-523.

[2] T. K. CAUGHEY AND J. ELLISON, *Existence, uniqueness and stability of a class of nonlinear partial differential equations*, J. Math. Anal. Appl., 51 (1975), pp. 1-32.

[3] J. CLEMENTS, *On the existence and uniqueness of solutions of the equation* $u_{tt} - \partial/\partial x_i\, \sigma_i(u_{x_i}) - \Delta_N u_t = f$, Canad. Math. Bull., 2 (1975), pp. 181-187.

[4] ———, *Existence theorem for a quasilinear evolution equation*, SIAM J. Appl. Math., 26 (1974), pp. 745-752.

[5] C. M. DAFERMOS, *The mixed initial boundary value problem for the equations of nonlinear one-dimensional viscoelasticity*, J. Differential Equations, 6 (1969), pp. 71-86.

[6] DANG DINH ANG AND A. PHAM NGOC DINH, *On the strongly damped wave equation*, to appear.

[7] N. DUNFORD AND J. SCHWARTZ, *Linear Operators. Part* II, Wiley Interscience, New York, 1963.

[8] A. FRIEDMAN, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.

[9] J. M. GREENBERG, R. C. MACCAMY AND V. J. MIZEL, *On the existence, uniqueness and stability of solutions of the equation* $\sigma'(u_x)u_{xx} + \lambda u_{xtx} = \rho_0 u_{tt}$, J. Math. Mech., 17 (1968), pp. 707-728.

[10] A. I. KOZHANOV, N. A. LARKIN AND N. N. JANENKO, *On a regularization of equations of variable type*, Dokl. Akad. Nauk. SSSR, 252 (1980), pp. 525-527. (In Russian.) Soviet Math. Dokl., 21 (1980), pp. 758-761. (In English.)

[11] V. LAKSHMIKANTHAM AND S. LEELA, *Differential and Integral Inequalities*, Vol. 1, Academic Press, New York-London, 1969.

[12] J. L. LIONS, *Quelques méthodes de résolution de problèmes aux limites non linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[13] A. PAZY, *A class of semi-linear equations of evolution*, Israel J. Math., 20 (1975), pp. 23-36.

[14] S. L. SOBOLEV, *Sur les Equations aux Dérivées Partielles Hyperboliques non Linéaires*, Edizioni Cremonese, Roma, 1961.

[15] G. F. WEBB, *Existence and asymptotic behavior for a strongly damped nonlinear wave equation*, Canad. J. Math., 32 (1980), pp. 631-643.

[16] Y. YAMADA, *Some remarks on the equation* $y_{tt} - \sigma(y_x)y_{xx} - y_{xtx} = f$, Osaka J. Math., 17 (1980), pp. 303-323.

[17] ———, *Quasilinear wave equations and related nonlinear evolution equations*, Nagoya Math. J., 84 (1981), pp. 31-83.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

007

DENSE SETS AND FAR FIELD PATTERNS FOR THE VECTOR HELMHOLTZ EQUATION UNDER TRANSMISSION BOUNDARY CONDITIONS*

PETER WILDE† AND ACHIM WILLERS†

Abstract. We study the set of far field patterns which are generated by entire incident fields scattered by a bounded penetrable obstacle. Necessary and sufficient conditions are given for the set to be dense in the set of all square integrable vector fields defined on the unit sphere and it is shown how these results can be generalized to Sobolev spaces and to classical function spaces.

Key words. far field pattern, scattering theory, Helmholtz equation, Fredholm integral equation

AMS(MOS) subject classifications. 78A45, 35J05, 45B05

1. Introduction. In the theory of direct problems in acoustic and electromagnetic scattering by bounded obstacles it is shown how the solution of a boundary value problem and the related far field corresponding to an incident field and to a given obstacle can be calculated. For further details see Colton and Kress [4].

Wilde [10], [11] has investigated the more general direct problem of electromagnetic scattering by a penetrable physical medium. To follow his considerations, let D_i be a bounded, simply connected domain in $\mathbb{R}^3$ with boundary S belonging to the class C^2 and define $D_e := \mathbb{R}^3 \setminus \overline{D_i}$. We assume that the normal vector n to S is directed into the exterior domain. To simplify notation, for any domain G with boundary ∂G of class C^2 we introduce the linear space of vector fields

$$F(G) := \{H : \bar{G} \to \mathbb{C}^3 \mid H \in C^2(G) \cap C(\bar{G}), \text{ curl } H, \text{ div } H \in C(\bar{G})\}.$$

Given an exterior wavenumber $\kappa_e > 0$, an interior wave number $\kappa_i \neq 0$ with $\operatorname{Im}(\kappa_i) \geqq 0$ and an entire solution H^i of the vector Helmholtz equation $\Delta H^i + \kappa_e^2 H^i = 0$ as incident field we consider the **magnetic transmission problem TH(m)** and the **electric transmission problem TH(e).**

Problem TH(m). Find two vector fields $H_e = H^i + H^s \in F(D_e)$ and $H_i \in F(D_i)$ satisfying the vector Helmholtz equations

$$\Delta H_e + \kappa_e^2 H_e = 0 \quad \text{in } D_e, \qquad \Delta H_i + \kappa_i^2 H_i = 0 \quad \text{in } D_i, \tag{1.1}$$

the *magnetic transmission conditions*

$$\begin{aligned} [n, H_e] \quad &- \quad [n, H_i] \quad = 0, \\ \operatorname{div} H_e \quad &- \quad \operatorname{div} H_i \quad = 0, \\ \beta_e[[\operatorname{curl} H_e, n], n] &- \beta_i[[\operatorname{curl} H_i, n], n] = 0, \\ \alpha_e(n, H_e) \quad &- \quad \alpha_i(n, H_i) \quad = 0 \quad \text{on } S \end{aligned} \tag{1.2}$$

and the *radiation condition*

$$[\operatorname{curl} H^s, \hat{x}] + \hat{x} \operatorname{div} H^s - \iota \kappa_e H^s = o(1/|x|), \qquad |x| \to \infty, \tag{1.3}$$

* Received by the editors October 28, 1985; accepted for publication (in revised form) January 23, 1987.
† Institut für Numerische und Angewandte Mathematik, Lotzestrasse 16-18, D-3400 Göttingen, Federal Republic of Germany.

uniformly for all directions $\hat{x} := x/|x|$, where $\iota := \sqrt{-1}$ is the imaginary unit and $\beta_e, \beta_i, \alpha_e, \alpha_i \in \mathbb{C}\backslash\{0\}$ are given complex numbers. H^s is called the scattered field.

By (a, b), $[a, b]$, and (a, b, c) we denote the scalar product, vector product, and triple scalar product of the vectors $a, b, c \in \mathbb{C}^3$, respectively.

In electromagnetic scattering the transmission parameters $\beta_e, \beta_i, \alpha_e$ and α_i are related to the dielectric constants, the permeability, the electric and magnetic conductivity, and to the frequency of the incoming wave (cf. Wilde [11]).

The problem obtained by replacing the magnetic transmission conditions (1.2) by the *electric transmission conditions*

$$\begin{aligned} \beta_e^{-1}[n, H_e] &- \beta_i^{-1}[n, H_i] = 0, \\ \alpha_e^{-1} \operatorname{div} H_e &- \alpha_i^{-1} \operatorname{div} H_i = 0, \\ [[\operatorname{curl} H_e, n], n] &- [[\operatorname{curl} H_i, n], n] = 0, \\ (n, H_e) &- (n, H_i) = 0 \quad \text{on } S \end{aligned} \tag{1.4}$$

is called **Problem TH(e).**

The corresponding inverse problem, i.e., to determine the shape of the penetrable scattering obstacle by measuring the scattered fields, is in many cases of more interest for physical application. Recently, Colton and Monk [6] presented an optimization scheme for solving the inverse scalar transmission problem. It is based on the decomposition of the space of all square integrable functions defined on the unit sphere into the set of all far field patterns corresponding to all incoming plane waves and a finite dimensional space of Herglotz kernels. This decomposition theorem, obtained by Kirsch [8], is equivalent to the assertion that, under certain circumstances, the class of far field patterns corresponding to a fixed scattering obstacle and all entire incident fields is dense in the set of all square integrable functions defined on the unit sphere.

Therefore, the classification of the class of far field patterns corresponding to the scattering of time harmonic incident fields by a penetrable bounded obstacle is one of the basic problems in inverse scattering theory for acoustics, electromagnetics and elastodynamics. Far field patterns for acoustic and electromagnetic scattering problems have been considered by many authors, e.g., Colton [1], Colton and Kirsch [2], [3], Colton and Kress [5], Kirsch [8] and Willers [12]. We shall carry out an analogous investigation for the transmission problems for the vector Helmholtz equations. Wilde [11] has shown that transmission problems for the vector Helmholtz equation are slight generalizations of transmission problems for the Maxwell equation. This may be the first step in solving the corresponding inverse problem.

As opposed to the authors mentioned above we shall only use integral equations in classical function spaces and avoid the theory of generalized boundary value problems.

In order to reduce the transmission problems to integral equations of the second kind we introduce the function spaces

$$C_T^{0,\alpha}(S) := \{a : S \to \mathbb{C}^3 | (a, n) = 0, a \in C^{0,\alpha}(S)\},$$

$$C_D^{0,\alpha}(S) := \{a \in C_T^{0,\alpha}(S) | \operatorname{Div} a \in C^{0,\alpha}(S)\},$$

$$P_\uparrow^{0,\alpha}(S) := C_D^{0,\alpha}(S) \times C^{0,\alpha}(S) \times C_T^{0,\alpha}(S) \times C^{0,\alpha}(S),$$

$$P_\downarrow^{0,\alpha}(S) := C_T^{0,\alpha}(S) \times C^{0,\alpha}(S) \times C_D^{0,\alpha}(S) \times C^{0,\alpha}(S),$$

and equip $C^{0,\alpha}(S)$, $C_T^{0,\alpha}(S)$ with the usual Hölder norms, $C_D^{0,\alpha}(S)$ with the norm $\|a\|_{1,\alpha} := \|a\|_\alpha + \|\operatorname{Div} a\|_\alpha$ and $P_\uparrow^{0,\alpha}(S)$, $P_\downarrow^{0,\alpha}(S)$ with the product norms. Furthermore,

we define the integral operators $\mathbf{M_m}, \mathbf{M_e}: \mathrm{P}_{\uparrow}^{0,\alpha}(S) \to P_{\uparrow}^{0,\alpha}(S)$, $\mathbf{M}'_{\mathbf{m}}, \mathbf{M}'_{\mathbf{e}}: P_{\downarrow}^{0,\alpha}(S) \to P_{\downarrow}^{0,\alpha}(S)$ by

$$\mathbf{M_m} := \begin{pmatrix} (I+L_e)\Omega_e + (I-L_i)\Omega_i & R_e - R_i \\ -\Gamma_e Q_e \Omega_e + \Gamma_i Q_i \Omega_i & \Gamma_e(I-L'_e) + \Gamma_i(I+L'_i) \end{pmatrix},$$

$$\mathbf{M}'_{\mathbf{m}} := \begin{pmatrix} \Omega_e(I+L'_e) + \Omega_i(I-L'_i) & -\Omega_e Q_e \Gamma_e + \Omega_i Q_i \Gamma_i \\ R_e - R_i & (I-L_e)\Gamma_e + (I+L_i)\Gamma_i \end{pmatrix},$$

$$\mathbf{M_e} := \begin{pmatrix} \Gamma_e^{-1}(I+L_e) + \Gamma_i^{-1}(I-L_i) & \Gamma_e^{-1} R_e \Omega_e - \Gamma_i^{-1} R_i \Omega_i \\ -Q_e + Q_i & (I-L'_e)\Omega_e + (I+L'_i)\Omega_i \end{pmatrix},$$

$$\mathbf{M}'_{\mathbf{e}} := \begin{pmatrix} (I+L'_e)\Gamma_e^{-1} + (I-L'_i)\Gamma_i^{-1} & -Q_e + Q_i \\ \Omega_e R_e \Gamma_e^{-1} - \Omega_i R_i \Gamma_i^{-1} & \Omega_e(I-L_e) + \Omega_i(I+L_i) \end{pmatrix},$$

where the 2×2 matrix operators L_e, L'_e, R_e, Q_e and L_i, L'_i, R_i, Q_i are the operators L, L', R, Q from [4, pp. 132, 138], [11] with respect to the wave numbers κ_e, κ_i, respectively, and where $\Gamma_e, \Omega_e, \Gamma_i, \Omega_i$ are given by

$$\Gamma_j := \begin{pmatrix} \beta_j I & 0 \\ 0 & \alpha_j I \end{pmatrix}, \qquad \Omega_j := \sigma_j \begin{pmatrix} \kappa_j^2 I & 0 \\ 0 & I \end{pmatrix}, \qquad j = e, i,$$

with

$$\sigma_j := \begin{cases} 1 & \text{if } \mathrm{Re}\,(\kappa_j) \geqq 0, \\ -1 & \text{otherwise,} \end{cases} \qquad j = e, i.$$

Note that the operators $\mathbf{M_m}, \mathbf{M}'_{\mathbf{m}}$ and the operators $\mathbf{M_e}, \mathbf{M}'_{\mathbf{e}}$ are adjoint with respect to the nondegenerate bilinear form

$$\langle \cdot, \cdot \rangle : P_{\uparrow}^{0,\alpha}(S) \times P_{\downarrow}^{0,\alpha}(S) \to \mathbb{C},$$

$$\langle \chi, \chi_0 \rangle := \int_S \{(a, a_0) + \lambda\lambda_0 + (b, b_0) + \delta\delta_0\}\, ds,$$

$\chi := (a, \lambda, b, \delta)^T \in P_{\uparrow}^{0,\alpha}(S)$, $\chi_0 := (a_0, \lambda_0, b_0, \delta_0)^T \in P_{\downarrow}^{0,\alpha}(S)$.

Before starting our analysis we summarize some results obtained by Wilde [11]. In these theorems $\Phi_i(x, y)$ and $\Phi_e(x, y)$ denote the fundamental solution

$$\Phi(x, y) := \frac{\exp(\iota\kappa|x-y|)}{4\pi|x-y|}, \qquad x, y \in \mathbb{R}^3, \quad x \neq y,$$

of the scalar Helmholtz equation with respect to the wave numbers κ_i and κ_e.

THEOREM 1.1. (i) *If* $\chi_m := (a, \lambda, b, \delta)^T \in P_{\uparrow}^{0,\alpha}(S)$ *is a solution of the integral equation*

$$\mathbf{M_m}\chi_m = f_m$$

with $f_m := 2(-[n, H^i], -\mathrm{div}\, H^i, \beta_e[[\mathrm{curl}\, H^i, n], n], \alpha_e(n, H^i))^T$, *then the fields*

$$H_i(x) := \mathrm{curl}_x \int_S a(y)\Phi_i(x, y)\, ds(y) - \int_S n(y)\lambda(y)\Phi_i(x, y)\, ds(y)$$

$$+ \sigma_i \kappa_i^2 \int_S [n(y), b(y)]\Phi_i(x, y)\, ds(y)$$

$$+ \sigma_i\, \mathrm{grad}_x \int_S \delta(y)\Phi_i(x, y)\, ds(y), \qquad x \in D_i,$$

$$H^s(x) := \operatorname{curl}_x \int_S a(y)\Phi_e(x,y)\,ds(y) - \int_S n(y)\lambda(y)\Phi_e(x,y)\,ds(y) + \sigma_e\kappa_e^2 \int_S [n(y), b(y)]\Phi_e(x,y)\,ds(y) + \sigma_e \operatorname{grad}_x \int_S \delta(y)\Phi_e(x,y)\,ds(y), \qquad x \in D_e, \tag{1.5}$$

are a solution of the magnetic transmission problem.

(ii) *If the homogeneous magnetic transmission problem admits only the trivial solution then the operators* $\mathbf{M_m}$ *and* $\mathbf{M'_m}$ *are invertible and the Problem* $TH(m)$ *is uniquely solvable.*

THEOREM 1.2. (i) *If* $\chi_e := (a, \lambda, b, \delta)^T \in P_\uparrow^{0,\alpha}(S)$ *is a solution of the integral equation*

$$\mathbf{M_e}\chi_e = f_e$$

with $f_e := 2(-\beta_e^{-1}[n, H^i], -\alpha_e^{-1} \operatorname{div} H^i, [[\operatorname{curl} H^i, n], n], (n, H^i))^T$, *then the fields*

$$H_i(x) := \operatorname{curl}_x \int_S a(y)\Phi_i(x,y)\,ds(y) - \int_S n(y)\lambda(y)\Phi_i(x,y)\,ds(y) + \sigma_i\kappa_i^2 \int_S [n(y), b(y)]\Phi_i(x,y)\,ds(y) + \sigma_i \operatorname{grad}_x \int_S \delta(y)\Phi_i(x,y)\,ds(y), \qquad x \in D_i,$$

$$H^s(x) := \operatorname{curl}_x \int_S a(y)\Phi_e(x,y)\,ds(y) - \int_S n(y)\lambda(y)\Phi_e(x,y)\,ds(y) + \sigma_e\kappa_e^2 \int_S [n(y), b(y)]\Phi_e(x,y)\,ds(y) + \sigma_e \operatorname{grad}_x \int_S \delta(y)\Phi_e(x,y)\,ds(y), \qquad x \in D_e, \tag{1.6}$$

are a solution of the electric transmission problem.

(ii) *If the homogeneous electric transmission problem admits only the trivial solution then the operators* $\mathbf{M_e}$ *and* $\mathbf{M'_e}$ *are invertible and the Problem* $TH(e)$ *is uniquely solvable.*

2. Dense sets and far field patterns. To start our investigations we prove an integral equation based on the representation theorem.

LEMMA 2.1. *Let the incident field* H^i *be an entire solution of the vector Helmholtz equation and let* H^s *be the scattered field under magnetic transmission conditions. Then the Cauchy data* $\chi_m^D := (\beta_e[[\operatorname{curl} H_e, n], n], \alpha_e(n, H_e), [n, H_e], \operatorname{div} H_e)^T$, *on* S, *of the exterior field satisfy the integral equation*

$$\mathbf{M'_e}\chi_m^D = f_m^D \tag{2.1}$$

with $f_m^D := 2([[\operatorname{curl} H^i, n], n], (n, H^i), \sigma_e\kappa_e^2[n, H^i], \sigma_e \operatorname{div} H^i)^T \in P_\downarrow^{0,\alpha}(S)$.

Proof. From the investigations of Colton and Kress [4] we obtain

$$\begin{pmatrix} [n, H_e] \\ \operatorname{div} H_e \end{pmatrix} - L_e \begin{pmatrix} [n, H_e] \\ \operatorname{div} H_e \end{pmatrix} + R_e \begin{pmatrix} [[\operatorname{curl} H_e, n], n] \\ (n, H_e) \end{pmatrix} = 2 \begin{pmatrix} [n, H^i] \\ \operatorname{div} H^i \end{pmatrix}, \tag{2.2}$$

$$-\begin{pmatrix}[n, H_i]\\ \operatorname{div} H_i\end{pmatrix} - L_i\begin{pmatrix}[n, H_i]\\ \operatorname{div} H_i\end{pmatrix} + R_i\begin{pmatrix}[[\operatorname{curl} H_i, n], n]\\ (n, H_i)\end{pmatrix} = 0, \tag{2.3}$$

$$\begin{pmatrix}[[\operatorname{curl} H_e, n], n]\\ (n, H_e)\end{pmatrix} - Q_e\begin{pmatrix}[n, H_e]\\ \operatorname{div} H_e\end{pmatrix} + L'_e\begin{pmatrix}[[\operatorname{curl} H_e, n], n]\\ (n, H_e)\end{pmatrix} = 2\begin{pmatrix}[[\operatorname{curl} H^i, n], n]\\ (n, H^i)\end{pmatrix}, \tag{2.4}$$

$$-\begin{pmatrix}[[\operatorname{curl} H_i, n], n]\\ (n, H_i)\end{pmatrix} - Q_i\begin{pmatrix}[n, H_i]\\ \operatorname{div} H_i\end{pmatrix} + L'_i\begin{pmatrix}[[\operatorname{curl} H_i, n], n]\\ (n, H_i)\end{pmatrix} = 0. \tag{2.5}$$

Inserting the transmission conditions into (2.3) and (2.5) we deduce

$$\begin{pmatrix}[n, H_e]\\ \operatorname{div} H_e\end{pmatrix} + L_i\begin{pmatrix}[n, H_e]\\ \operatorname{div} H_e\end{pmatrix} - R_i\Gamma_i^{-1}\Gamma_e\begin{pmatrix}[[\operatorname{curl} H_e, n], n]\\ (n, H_e)\end{pmatrix} = 0, \tag{2.6}$$

$$\Gamma_i^{-1}\Gamma_e\begin{pmatrix}[[\operatorname{curl} H_e, n], n]\\ (n, H_e)\end{pmatrix} + Q_i\begin{pmatrix}[n, H_e]\\ \operatorname{div} H_e\end{pmatrix} - L'_i\Gamma_i^{-1}\Gamma_e\begin{pmatrix}[[\operatorname{curl} H_e, n], n]\\ (n, H_e)\end{pmatrix} = 0. \tag{2.7}$$

Adding (2.7) and (2.4) we obtain the first part of (2.1). To prove the second part we have to multiply (2.2) by Ω_e and (2.6) by Ω_i and add the equations. □

Now we write some definitions that we need in order to formulate the main theorem of this paper.

DEFINITION 2.2. κ_e is called an *eigenvalue of the magnetic transmission problem* if there exist nontrivial vector fields $G, W \in F(D_i)$ satisfying the vector Helmholtz equations

$$\Delta W + \kappa_e^2 W = 0 \quad \text{in } D_i, \qquad \Delta G + \kappa_i^2 G = 0 \quad \text{in } D_i, \tag{2.8}$$

and the boundary conditions

$$\begin{aligned} [n, W] \quad &- \quad [n, G] \quad = 0, \\ \operatorname{div} W \quad &- \quad \operatorname{div} G \quad = 0, \\ \beta_e[[\operatorname{curl} W, n], n] &- \beta_i[[\operatorname{curl} G, n], n] = 0, \\ \alpha_e(n, W) \quad &- \quad \alpha_i(n, G) \quad = 0 \qquad \text{on } S. \end{aligned} \tag{2.9}$$

The field W is called the *corresponding eigenfunction.* κ_e is called an *eigenvalue of the electric transmission problem* if we have

$$\begin{aligned} \beta_e^{-1}[n, W] \quad &- \quad \beta_i^{-1}[n, G] \quad = 0, \\ \alpha_e^{-1} \operatorname{div} W \quad &- \quad \alpha_i^{-1} \operatorname{div} G \quad = 0, \\ [[\operatorname{curl} W, n], n] &- [[\operatorname{curl} G, n], n] = 0, \\ (n, W) \quad &- \quad (n, G) \quad = 0 \qquad \text{on } S \end{aligned} \tag{2.10}$$

instead of (2.9).

Such eigenvalues may exist. This is shown by the following example.

Example 2.3. In the special case where S is the unit sphere in $\mathbb{R}^3$ we define

$$G(x) := \frac{\sin \kappa_e}{\kappa_i^3 \kappa_e} \operatorname{grad}\left(\frac{\sin \kappa_i |x|}{|x|}\right), \qquad x \in D_i,$$

$$W(x) := \frac{\sin \kappa_i}{\kappa_e^3 \kappa_i} \operatorname{grad}\left(\frac{\sin \kappa_e |x|}{|x|}\right), \qquad x \in D_i,$$

where κ_e and κ_i are different solutions of the equation $\tan t = t$.

Now $\{G, W\}$ is a solution of (2.8) and from

$$G(x) = \frac{\sin \kappa_e}{\kappa_e}\left(\frac{\kappa_i \cos \kappa_i|x|}{|x|} - \frac{\sin \kappa_i|x|}{|x|^2}\right)\hat{x},$$

$$W(x) = \frac{\sin \kappa_i}{\kappa_i}\left(\frac{\kappa_e \cos \kappa_e|x|}{|x|} - \frac{\sin \kappa_e|x|}{|x|^2}\right)\hat{x},$$

we easily see that the homogeneous boundary conditions (2.9) are satisfied.

On the other hand it can be shown that under certain assumption on κ_e, κ_i, α_e, α_i, β_e, and β_i the boundary value problem (2.8), (2.10) admits only the trivial solution.

Example 2.4. Let the following relations be fulfilled:

$$\gamma := \frac{\alpha_e}{\alpha_i} = \frac{\overline{\beta_e}}{\overline{\beta_i}} \quad \text{with } \gamma \notin \mathbb{R}, \quad \kappa_e > 0 \quad \text{and Re}\,(\kappa_i) = 0.$$

Then the first vector Green's theorem yields

$$\begin{aligned}\int_D \{-\kappa_e^2|W|^2 + |\text{curl } W|^2 + |\text{div } W|^2\}\, dx &= \int_S \{(n, \bar{W}, \text{curl } W) + (n, \bar{W})\, \text{div } W\}\, ds \\ &= \bar{\gamma} \int_S \{(n, \bar{G}, \text{curl } G) + (n, \bar{G})\, \text{div } G\}\, ds \\ &= \bar{\gamma} \int_D \{-\kappa_i^2|G|^2 + |\text{curl } G|^2 + |\text{div } G|^2\}\, dx\end{aligned}$$

and we obtain $G = W = 0$.

DEFINITION 2.5. An entire solution W of the vector Helmholtz equation is called a *Herglotz field* if

$$\lim_{r\to\infty} \frac{1}{r} \int_{|x|\leq r} |W(x)|^2\, dx < \infty.$$

Let $\mathbf{Y}$ be the set of all square integrable vector fields defined on the unit sphere. Applying the results of Hartman and Wilcox [7] to each coordinate of W we obtain Lemma 2.6.

LEMMA 2.6. *An entire solution W of the vector Helmholtz equation $\Delta W + \kappa^2 W = 0$ is a Herglotz field if and only if there exists a vector field $w \in \mathbf{Y}$ such that W has the representation*

$$W(x) = \int_{|\hat{y}|=1} w(\hat{y})\, e^{i\kappa(x,\hat{y})}\, ds(\hat{y}), \qquad x \in \mathbb{R}^3.$$

Furthermore, w is uniquely determined by W.

By using the representation theorem for solutions of the vector Helmholtz equation Colton and Kress [4] have shown the relation

$$H^s(x) = \frac{e^{i\kappa_e|x|}}{4\pi|x|} F(\hat{x}) + O\left(\frac{1}{|x|^2}\right), \qquad |x| \to \infty,$$

uniformly for all directions $\hat{x} \in \mathbb{R}^3$, where the vector field F is defined by

$$F(\hat{x}) := \iota\kappa_e \int_S [\hat{x}, [n(y), H_e(y)]]\, e^{-\iota\kappa_e(\hat{x},y)}\, ds(y)$$
$$- \int_S n(y)\, \operatorname{div} H_e(y)\, e^{-\iota\kappa_e(\hat{x},y)}\, ds(y)$$
$$- \int_S [\operatorname{curl} H_e(y), n(y)]\, e^{-\iota\kappa_e(\hat{x},y)}\, ds(y)$$
$$- \iota\kappa_e \int_S \hat{x}(n(y), H_e(y))\, e^{-\iota\kappa_e(\hat{x},y)}\, ds(y), \qquad \hat{x} \in \Gamma,$$

where Γ denotes the unit sphere in $\mathbb{R}^3$. Furthermore, they have shown that $F = 0$ implies $H^s = 0$. The vector field F is called the *far field pattern* of the scattered field H^s.

Now we are able to prove a theorem on the set $\mathbf{F}^{\mathbf{m}}_{\kappa_e}$ of far field patterns corresponding to all entire incident fields under magnetic transmission conditions.

THEOREM 2.7. *If the homogeneous magnetic transmission problem has only the trivial solution, then we have*:

If κ_e is not an eigenvalue of the electric transmission problem, then $\mathbf{F}^{\mathbf{m}}_{\kappa_e}$ is dense in $\mathbf{Y}$. If κ_e is an eigenvalue of the electric transmission problem, then $\mathbf{F}^{\mathbf{m}}_{\kappa_e}$ is dense in $\mathbf{Y}$ if and only if none of the corresponding eigenfunctions is a Herglotz field.

Proof. Let $\mathbf{F}^{\mathbf{m}}_{\kappa_e}$ be dense in $\mathbf{Y}$ and let W be a Herglotz field which is an eigenfunction in D_i. W has a representation of the form

$$W(x) := \int_{|\hat{y}|=1} \overline{w(\hat{y})}\, e^{-\iota\kappa_e(x,\hat{y})}\, ds(\hat{y}), \qquad x \in \mathbb{R}^3, \tag{2.11}$$

for some $w \in \mathbf{Y}$.

If $F \in \mathbf{F}^{\mathbf{m}}_{\kappa_e}$ is a given far field pattern and H_e, H_i is the corresponding solution of the magnetic transmission problem in D_e, D_i, respectively, then from the second vector Green's theorem we obtain

$$\int_{|\hat{x}|=1} (\overline{w(\hat{x})}, F(\hat{x}))\, ds(\hat{x})$$
$$= \int_S (n, H_e, \operatorname{curl} W)\, ds + \int_S (n, H_e) \operatorname{div} W\, ds$$
$$- \int_S (n, W, \operatorname{curl} H_e)\, ds - \int_S (n, W) \operatorname{div} H_e\, ds$$
$$= \int_S (n, H_i, \operatorname{curl} G)\, ds + \int_S (n, H_i) \operatorname{div} G\, ds$$
$$- \int_S (n, G, \operatorname{curl} H_i)\, ds - \int_S (n, G) \operatorname{div} H_i\, ds = 0.$$

Now the density of $\mathbf{F}^{\mathbf{m}}_{\kappa_e}$ in $\mathbf{Y}$ implies $w = 0$ and therefore $W = 0$.

Conversely, suppose that there exists a vector field $w \in \mathbf{Y}$ such that

$$\int_{|\hat{x}|=1} (\overline{w(\hat{x})}, F(\hat{x}))\, ds(\hat{x}) = 0$$

for all $F \in \mathbf{F}^{\mathbf{m}}_{\kappa_e}$.

Defining W as in (2.11) and using the integral equation (2.1), we obtain

$$
\begin{aligned}
0 &= \int_{|\hat{x}|=1} \overline{(w(\hat{x}), F(\hat{x}))}\, ds(\hat{x}) \\
&= \int_S (n, H_e, \operatorname{curl} W)\, ds + \int_S (n, H_e) \operatorname{div} W ds - \int_S (n, W, \operatorname{curl} H_e)\, ds \\
&\quad - \int_S (n, W) \operatorname{div} H_e\, ds \\
&= \left\langle \begin{pmatrix} \beta_e[[\operatorname{curl} H_e, n], n] \\ \alpha_e(n, H_e) \\ [n, H_e] \\ \operatorname{div} H_e \end{pmatrix}, \begin{pmatrix} \beta_e^{-1}[n, W] \\ \alpha_e^{-1} \operatorname{div} W \\ -[[\operatorname{curl} W, n], n] \\ -(n, W) \end{pmatrix} \right\rangle \qquad (2.12) \\
&= 2 \left\langle (\mathbf{M}_e')^{-1} \begin{pmatrix} [[\operatorname{curl} H^i, n], n] \\ (n, H^i) \\ \sigma_e \kappa_e^2 [n, H^i] \\ \sigma_e \operatorname{div} H^i \end{pmatrix}, \begin{pmatrix} \beta_e^{-1}[n, W] \\ \alpha_e^{-1} \operatorname{div} W \\ -[[\operatorname{curl} W, n], n] \\ -(n, W) \end{pmatrix} \right\rangle \\
&= \langle (\mathbf{M}_e')^{-1} f_m^D, g \rangle = \langle f_m^D, \mathbf{M}_e^{-1} g \rangle,
\end{aligned}
$$

where g is given by $g := (\beta_e^{-1}[n, W], \alpha_e^{-1} \operatorname{div} W, -[[\operatorname{curl} W, n], n], -(n, W))^T$ and f_m^D was defined in (2.1). With $\chi_e := (a, \lambda, b, \delta)^T := -2\mathbf{M}_e^{-1} g$ we define vector fields G^s and G_i as in (1.6). We observe that G_i and $G_e := G^s + W$ form a solution of *Problem TH*(e). Defining

$$
v_k^m(x) := h_k^{(1)}(\kappa_e |x|)\, Y_k^m(\hat{x}), \qquad u_k^m(x) := j_k(\kappa_e |x|)\, Y_k^m(\hat{x}),
$$

where $h_k^{(1)}$ denote the spherical Hankel function of the first kind of order k, j_k the spherical Bessel function of order k and Y_k^m the spherical harmonics of order k, $k = 0, 1, 2, \cdots$, $m = -k, \cdots, k$, and using the expansion $\Phi_e(x, y) = \iota\kappa_e \sum_{k=0}^{\infty} \sum_{m=-k}^{k} v_k^m(x) u_k^m(y)$, $|y| < |x|$, for all sufficiently large $|x|$ and every unit vector c, we obtain

$$
\begin{aligned}
(c, G^s(x)) &= \int_S (a(y), \operatorname{curl}_y \{c\Phi_e(x, y)\})\, ds(y) \\
&\quad - \int_S (c, n(y))\lambda(y)\Phi_e(x, y)\, ds(y) \\
&\quad + \sigma_e \kappa_e^2 \int_S (c, n(y), b(y))\Phi_e(x, y)\, ds(y) \\
&\quad - \sigma_e \int_S \delta(y) \operatorname{div}_y \{c\Phi_e(x, y)\}\, ds(y) \\
&= -\iota\kappa_e \sum_{k=0}^{\infty} \sum_{m=-k}^{k} v_k^m(x) 2\langle R_k^m, \mathbf{M}_e^{-1} g \rangle,
\end{aligned}
$$

where $R_k^m := ([[\operatorname{curl} E_k^m, n], n], (n, E_k^m), \sigma_e \kappa_e^2 [n, E_k^m], \sigma_e \operatorname{div} E_k^m)^T$ and $E_k^m := c u_k^m$.

Since in (2.12) we can especially choose $H^i = E_k^m$ as incident fields, it follows from the analyticity of G^s that G^s vanishes in D_e. Thus W is an eigenfunction of the electric transmission problem and the proof is completed by the second part of Lemma 2.6. □

Denoting with $\mathbf{F}^{e}_{\kappa_e}$ the set of far field patterns corresponding to all entire incident fields under electric transmission conditions we are able to prove the following theorem in an analogous way.

THEOREM 2.8. *If the homogeneous electric transmission problem has only the trivial solution, then we have*:

If κ_e is not an eigenvalue of the magnetic transmission problem, then $\mathbf{F}^{e}_{\kappa_e}$ is dense in $\mathbf{Y}$. If κ_e is an eigenvalue of the magnetic transmission problem, then $\mathbf{F}^{e}_{\kappa_e}$ is dense in $\mathbf{Y}$ if and only if none of the corresponding eigenfunctions is a Herglotz field.

Remark 2.9. Sufficient conditions for Problem TH(m) and Problem TH(e) to have at most one solution are given in [10], [11].

3. Far field patterns in Sobolev spaces. In this section we shall briefly discuss the denseness of far field patterns in Sobolev spaces.

Analogous to Nečas [9] for $r \in (0, 1)$ we define the Sobolev space $\mathbf{Y}^r$ to be the completion of the set of continuous differentiable vector fields defined on the unit sphere with respect to the inner product

$$(v, w)_r := \int_{|\hat{x}|=1} (v(\hat{x}), \overline{w(\hat{x})})\, ds(\hat{x}) + \int_{|\hat{z}|=1} \int_{|\hat{x}|=1} \frac{(v(\hat{x}) - v(\hat{z}), \overline{w(\hat{x})} - \overline{w(\hat{z})})}{|\hat{x} - \hat{z}|^{2r+2}}\, ds(\hat{x})\, ds(\hat{z}).$$

Now, motivated by Lemma 2.6, we generalize Definition 2.5 in the following way.

DEFINITION 3.1. An entire solution W of the vector Helmholtz equation $\Delta W + \kappa^2 W = 0$ is called a *Herglotz field of index r* if there exists a vector field $w \in \mathbf{Y}^r$ such that W has the representation

$$W(x) = \sum_{j=1}^{3} (w, e_j\, e^{-\iota\kappa(x, \cdot)})_r\, e_j, \qquad x \in \mathbb{R}^3,$$

where $e_j, j = 1, 2, 3$, denote the Cartesian unit coordinate vectors.

Using this definition we can state a theorem analogous to Theorem 2.7.

THEOREM 3.2. *If the homogeneous magnetic transmission problem has only the trivial solution, then we have*:

If κ_e is not an eigenvalue of the electric transmission problem, then $\mathbf{F}^{m}_{\kappa_e}$ is dense in $\mathbf{Y}^r$. If κ_e is an eigenvalue of the electric transmission problem, then $\mathbf{F}^{m}_{\kappa_e}$ is dense in $\mathbf{Y}^r$ if and only if none of the corresponding eigenfunctions is a Herglotz field of index r.

Proof. Let $F \in \mathbf{F}^{m}_{\kappa_e}$ be a given far field pattern and H_e the corresponding solution of the magnetic transmission problem in D_e. If W has the representation

$$W(x) = \sum_{j=1}^{3} (\bar{w}, e_j\, e^{\iota\kappa(x, \cdot)})_r\, e_j, \qquad x \in \mathbb{R}^3,$$

for some $w \in \mathbf{Y}^r$, then by interchanging the order of integration we obtain

$$(F, w)_r = \int_S (n, H_e, \operatorname{curl} W)\, ds + \int_S (n, H_e) \operatorname{div} W\, ds - \int_S (n, W, \operatorname{curl} H_e)\, ds - \int_S (n, W) \operatorname{div} H_e\, ds.$$

The rest of the proof is exactly the same as in Theorem 2.7. □

Analogous considerations are possible as well for the electric transmission problem as for arbitrary $r > 0$. Furthermore, using Sobolev's embedding theorems, we are able to consider the denseness of far field patterns in the sets of continuous and Hölder continuous functions, also.

Finally we emphasize that we had not to require the boundary S to be smoother than C^2 throughout this paper.

REFERENCES

[1] D. COLTON, *Far field patterns for the impedance boundary value problem in acoustic scattering*, Appl. Anal., 16 (1983), pp. 131-139.

[2] D. COLTON AND A. KIRSCH, *Dense sets and far field patterns in acoustic wave propagation*, this Journal, 15 (1984), pp. 996-1006.

[3] ———, *Dense sets and far field patterns for the transmission problem*, in Proc. Conference on Classical Scattering Theory, G. Roach, ed., Shiva, Glasgow, Scotland 1984.

[4] D. COLTON AND R. KRESS, *Integral Equation Methods in Scattering Theory*, Wiley-Interscience, New York, 1983.

[5] ———, *Dense sets and far field patterns in electromagnetic wave propagation*, this Journal, 16 (1985), pp. 1049-1060.

[6] D. COLTON AND P. MONK, *The Inverse Scattering Problem for Time-Harmonic Acoustic Waves in a Penetrable Medium*, University of Delaware, Newark, DE, 1986, preprint.

[7] P. HARTMAN AND C. WILCOX, *On solutions of the Helmholtz equation in exterior domains*, Math. Z., 75 (1961), pp. 228-255.

[8] A. KIRSCH, *Generalized boundary value- and control problems for the Helmholtz equation*, Habilitation thesis, Göttingen, West Germany, 1984.

[9] J. NEČAS, *Les méthodes directes en théorie des équations elliptiques*. Masson, Paris, 1967.

[10] P. WILDE, *Über Transmissionsprobleme bei der vektoriellen Helmholtzgleichung*, Dissertation, Universität Göttingen, West Germany, 1985.

[11] ———, *Transmission problems for the vector Helmholtz equation*, Proc. Royal Soc. Edinburgh, Sect. A, 105A (1987), pp. 61-76.

[12] A. WILLERS, *Integral equations methods for the Helmholtz equation in disturbed half-spaces*, Math. Meth. Appl. Sci., to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

008

SOME REMARKS ABOUT THE MORSE LEMMA IN INFINITE DIMENSION*

MICHEL CROUZEIX†, GIUSEPPE GEYMONAT‡, AND GENEVIÈVE RAUGEL§

Abstract. We prove a Morse lemma for functionals that are of class C^2 (in a weak sense) on a Banach space. We also give a splitting lemma for C^2 functionals.

Key words. Morse lemma, splitting lemma, critical point, minima

AMS(MOS) subject classification. 58E05

1. Introduction. In this paper we prove a Morse lemma that, in some ways, improves Golubitsky and Marsden's theorem [5] as well as that of Tromba [12]. Indeed we relax the C^3-regularity condition required in [5] for the functional; under assumptions that are slightly different from those of Golubitsky and Marsden, we show that, when the functional is of class C^1 and admits a gradient which is "generator-differentiable" in the sense of Hugues and Marsden [6], it is equal to its quadratic part up to a homeomorphism, which is "C^1-generator-differentiable" in the sense of [6]. Furthermore we give a "splitting lemma" which is useful in the study of bifurcation problems and is true for C^2 functionals; it also contains a generalization of the "generalized Morse lemma" of Mawhin and Willem [10]. Actually, we have two things in mind here: the first is to complete the papers of Golubitsky and Marsden [5] and of Buchner, Marsden and Schecter [1]; the second is to give results that are applicable to elliptic variational problems (which have only few regularity properties) and also, after some minor modifications, to their approximate (or discrete) versions. In order to easily extend our results to approximate problems, we have used elementary and rough enough tools in our proofs; for instance, *the change of variables in our Morse lemma is given by an explicit formula*, which can be easily adapted to the approximate case. As the generalization to the approximate case is somewhat technical, we shall not give it here. (For the statement and the proof of the approximate splitting lemma, we refer the reader to [11].)

An outline of the paper is as follows. In § 2 we state and prove our generalization of the Morse lemma (see Theorem 2.1) as well as parametrized versions of it; we also give an example of application. Section 3 is devoted to the proof of the splitting lemma; moreover, we describe the relationship of the splitting lemma with the Morse lemma.

2. The Morse lemma.

2.1. Let X be a (real) Banach space equipped with the norm $\|\cdot\|_X$ and let H be a Hilbert space such that X is included in H with a continuous and dense imbedding. We denote by $\langle\cdot,\cdot\rangle$ the inner product of H and by $\|\cdot\|_H$ the associated norm. We identify the space H with its dual space so that we obtain the continuous imbeddings: $X \hookrightarrow H \hookrightarrow X'$. We also introduce a functional f on X as well as a continuous linear operator T from X into X, (which will be the second derivative of f at the point 0).

We assume that T and f satisfy the following hypotheses:

(H.1) T is an isomorphism of X onto X and is a symmetric operator with respect

* Received by the editors December 30, 1985; accepted for publication (in revised form) April 6, 1987.

† Université de Rennes, Mathématiques-IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France.

‡ Ecole Normale Supérieure de Cachan, Laboratoire de Mécanique et Technologie, 61, avenue du Président Wilson, 94230 Cachan, France.

§ Ecole Polytechnique Centre de Mathématiques Appliquées, Unité de Recherche Associée au C.N.R.S.-756, 91128 Palaiseau Cedex, France.

to the inner product $\langle \cdot, \cdot \rangle$, i.e.,

$$\forall x \in X \quad \forall y \in X \quad \langle Tx, y \rangle = \langle x, Ty \rangle; \tag{2.1}$$

(H.2) there exists a neighbourhood $\mathcal{V}$ of 0 in X such that $f \in C^1(\mathcal{V}; \mathbb{R})$ and

$$f(0) = 0, \qquad Df(0) = 0; \tag{2.2}$$

(H.3) $Df(x) \in C^0(\mathcal{V}; H)$ and converges to Tx in the following way:

$$\forall x \in \mathcal{V} \quad \|Df(x) - Tx\|_H = \|x\|_H \varepsilon_1(x), \tag{2.3}$$

where

$$\lim_{\|x\|_X \to 0} \varepsilon_1(x) = 0. \tag{2.4}$$

(In order to simplify our notation, we set $\varepsilon_1(0) = 0$.)

The assumption (H.3) actually means that f admits a gradient relative to $\langle \cdot, \cdot \rangle$ that takes its values in H and is generator-differentiable at 0 in the sense of Hugues and Marsden [6]. Let us recall that a mapping F from an open set $\mathcal{U}$ in X into H is "generator-differentiable" if, for any x in $\mathcal{U}$, there exists a continuous linear operator $DF(x)$ from X into H such that

$$\frac{\|F(x+h) - F(x) - DF(x) \cdot h\|_H}{\|h\|_H} \to 0 \quad \text{as } \|h\|_X \to 0.$$

If X and Y are two normed vector spaces, let us denote by $\mathcal{L}(X; Y)$ the space of all continuous linear operators from X into Y and by $\mathcal{L}(X)$ the space $\mathcal{L}(X; X)$.

Now we may state the following main result.

THEOREM 2.1. *Assume that the hypothesis* (H.1) *is satisfied. Then the assumptions* (H.2) *and* (H.3) *hold if and only if there exists a homeomorphism* Φ *of a neighbourhood* $\tilde{\mathcal{V}}$ *of* 0 *in* X *onto a neighbourhood* $\mathcal{W}$ *of* 0 *in* X *satisfying*:

$$\forall x \in \tilde{\mathcal{V}} \quad f(x) = \tfrac{1}{2}\langle T\Phi(x), \Phi(x) \rangle, \tag{2.5}$$

$$\Phi \in C^0(\tilde{\mathcal{V}}; X) \cap C^1(\tilde{\mathcal{V}} - \{0\}; X), \tag{2.6}$$

$$\Phi(0) = 0, \qquad D\Phi(0) = Id_H, \tag{2.7}$$

$$D\Phi \in C^0(\tilde{\mathcal{V}}; \mathcal{L}(H)). \tag{2.8}$$

In Remark 2.3, we shall show that, if the assumptions (H.2) and (H.3) hold, the homeomorphism Φ obtained above is actually C^1-generator-differentiable (let us recall that a mapping F from an open set $\mathcal{U}$ into H is C^1-generator-differentiable if it is generator-differentiable and $DF(x)$ belongs to $C^0(\mathcal{U}; \mathcal{L}(X; H))$).

Theorem 2.1 together with the implicit function theorem give us the following corollary, at once.

COROLLARY 2.2. *Assume that the hypothesis* (H.1) *is true and that* $X = H$. *Then the assumptions* (H.2) *and* (H.3) *hold if and only if there exists a* C^1*-diffeomorphism* Φ *of a neighbourhood* $\tilde{\mathcal{V}}$ *of* 0 *in* X *onto a neighbourhood* $\mathcal{W}$ *of* 0 *in* X *satisfying the properties* (2.5) *and* (2.7).

Let us point out that the mapping Φ, given in Theorem 2.1, is not the same as that of Golubitsky and Marsden.

Now let us replace the hypothesis (H.3) by the following hypothesis:

(H.3 bis) $Df(x) \in C^0(\mathcal{V}; H)$ and converges to Tx in the following way:

(2.3 bis) $$\forall x \in \mathcal{V} \quad \begin{cases} |f(x) - \frac{1}{2}\langle Tx, x\rangle| = \|x\|_H^3 \eta_2(x), \\ \|Df(x) - Tx\|_{X'} = \|x\|_H^2 \varepsilon_2(x), \end{cases}$$

where

(2.4 bis) $$\forall x \in \mathcal{V} \quad \eta_2(x) + \varepsilon_2(x) \leqq C,$$

C being a positive constant.

THEOREM 2.3. *Assume that the hypotheses* (H.1), (H.2) *and* (H.3 bis) *hold. Then there exists a* C^1*-diffeomorphism* Φ *of a neighbourhood* $\tilde{\mathcal{V}}$ *of* 0 *in* X *onto a neighbourhood* $\mathcal{W}$ *of* 0 *in* X *such that* (2.5) *and* (2.7) *are satisfied.*

Let us make a few remarks before proving Theorem 2.1 and Theorem 2.3.

Remark 2.1. Let us recall that Golubitsky and Marsden [5] proved that, if the assumptions

(A.1) there exists a neighbourhood $\mathcal{V}$ of 0 in X such that $f \in C^3(\mathcal{V}; \mathbb{R})$ and f admits a gradient ∇f relative to $\langle \cdot, \cdot \rangle$ that belongs to $C^2(\mathcal{V}; X)$, i.e.,

(2.9) $$\forall x \in \mathcal{V} \quad \forall x' \in X \quad Df(x)x' = \langle \nabla f(x), x' \rangle \quad \text{with } \nabla f \in C^2(\mathcal{V}; X)$$

and

(A.2) $f(0) = 0$, $Df(0) = 0$, $D^2 f(0) = T$, where T is an isomorphism of X onto X,

hold, there exists a C^1-diffeomorphism ψ of a neighbourhood N of 0 in X onto a neighbourhood M of 0 in X, satisfying:

(2.10) $$\forall x \in N \quad f(x) = \tfrac{1}{2}\langle T\psi(x), \psi(x)\rangle.$$

The conditions required for the existence of a homeomorphism Φ of a neighbourhood $\tilde{\mathcal{V}}$ of 0 in X onto a neighbourhood $\mathcal{W}$ of 0 in X satisfying (2.5) and (2.7) are weaker than the conditions (A.1) and (A.2); in particular, we need C^2-regularity assumptions at most. Indeed, using Lemma 2.4 below, the reader will at once see that the hypotheses (H.1)-(H.3) are satisfied if the assumption (A.2) holds and if f and Df belong to $C^2(\mathcal{V}; \mathbb{R})$ and $C^1(\mathcal{V}; X)$, respectively.

As stated in Theorem 2.3, the homeomorphism Φ becomes a C^1-diffeomorphism when we replace the hypothesis (H.3) by the hypothesis (H.3 bis). Note that the condition (H.3 bis) is a kind of three times differentiability assumption at 0. But the assumptions (A.1) and (A.2) are neither weaker, nor stronger than the conditions (H.1), (H.2) and (H.3 bis). For instance, if $X = C^0([0, 1])$, $H = L^2(0, 1)$ (provided with its usual inner product) and $f(x) = \int_0^1 [(x(t))^2 + (x(t))^3]\, dt$, the assumptions (A.1) and (A.2) are clearly satisfied, while the condition (H.3 bis) fails; conversely, if we keep the same spaces X and H and if $f(x) = \int_0^1 (x(t))^2\, dt + [\int_0^1 g(t)x(t)\, dt]^3$, where $g \in L^2(0, 1)$, but $g \notin C^0([0, 1])$, the conditions (H.1), (H.2) and (H.3 bis) hold, but (A.1) is not satisfied.

Finally, let us remark that the hypothesis (H.3 bis) holds if, for instance,

(2.11) $$\|Df(x) - Tx\|_H = \|x\|_H^2 \tilde{\varepsilon}_2(x),$$

where, for any x in $\mathcal{V}$, $\tilde{\varepsilon}_2(x) \leqq C$.

Now let us state an auxiliary result which will be useful in the proof of Theorem 2.1.

LEMMA 2.4. *If A is a continuous linear operator from X into X and satisfies*

$$\forall x \in X \quad \forall y \in X \quad \langle Ax, y\rangle = \langle Ay, x\rangle, \tag{2.12}$$

it extends into a continuous linear operator from H into H, still denoted by A, and the inequalities

$$\|A\|_{\mathscr{L}(H)} \leqq \rho_X(A) \leqq \|A\|_{\mathscr{L}(X)}, \tag{2.13}$$

hold, where $\rho_X(A)$ is the spectral radius of A considered as an operator from X into X.

The proof of Lemma 2.4 can be found in [4, Vol. I, p. 346] or in [8], for instance.

Remark 2.2. Lemma 2.4 implies in particular that, if T is an isomorphism of X onto X, it can be extended into an isomorphism of H onto H.

Proof of Theorem 2.1. (1) Assume that there exists a homeomorphism Φ of $\tilde{\mathscr{V}}$ onto $\mathscr{W}$ satisfying (2.5)–(2.8), and that (H.1) holds. Without any loss of generality we may assume that $\tilde{\mathscr{V}}$ is path-connected. Using the properties (2.6) to (2.8) together with a Taylor formula, we at once show that

$$\forall x \in \tilde{\mathscr{V}} \quad \|\Phi(x)\|_H \leqq C\|x\|_H, \tag{2.14}$$

where $C > 0$ is a constant independent of h, and that,

$$\forall x \in \tilde{\mathscr{V}} \quad \|\Phi(x) - x\|_H = \|x\|_H \varepsilon_3(x), \tag{2.15}$$

where

$$\lim_{\|x\|_X \to 0} \varepsilon_3(x) = 0. \tag{2.16}$$

Owing to the property (2.5), we may write

$$Df(x) = (D\Phi(x))^t T\Phi(x), \tag{2.17}$$

which implies, due to (2.8), that $f \in C^1(\tilde{\mathscr{V}}; \mathbb{R})$ and $Df \in C^0(\tilde{\mathscr{V}}; H)$. From (2.17) we also derive:

$$Df(x) - Tx = (D\Phi(x) - Id_H)^t T\Phi(x) + T(\Phi(x) - x). \tag{2.18}$$

Equality (2.18) together with the properties (2.8), (2.15) and (2.16) give us the conditions (2.3) and (2.4).

(2) Now assume that the hypotheses (H.1) to (H.3) hold. First of all, let us remark that, thanks to the hypothesis (H.3), we have:

$$\begin{aligned} |f(x) - \tfrac{1}{2}\langle Tx, x\rangle| &= \left| \int_0^1 \langle Df(sx) - T(sx), x\rangle \, ds \right| \\ &\leqq \|x\|_H^2 \int_0^1 s\varepsilon_1(sx) \, ds, \end{aligned}$$

or

$$|f(x) - \tfrac{1}{2}\langle Tx, x\rangle| \leqq \|x\|_H^2 \varepsilon_4(x) \tag{2.19}$$

where

$$\lim_{\|x\|_X \to 0} \varepsilon_4(x) = 0 \quad \text{and} \quad \varepsilon_4(0) = 0. \tag{2.20}$$

As by Lemma 2.4 and the hypothesis (H.1) (see Remark 2.2), T^{-1} belongs to $\mathscr{L}(H)$, we infer from (2.19) and (2.20) that there exists a (convex) neighbourhood V of 0 in X such that

$$\forall x \in V \quad |(2f(x) - \langle Tx, x\rangle)\langle T^{-1}x, x\rangle| \leqq \tfrac{1}{2}\|x\|_H^4, \tag{2.21}$$

which allows us to define the function α in $C^0(V;\mathbb{R})\cap C^1(V-\{0\};\mathbb{R})$ by

$$\alpha(x)=\frac{2f(x)-\langle Tx,x\rangle}{\sqrt{\|x\|_H^4+(2f(x)-\langle Tx,x\rangle)\langle T^{-1}x,x\rangle}+\|x\|_H^2}\quad\text{for } x\neq 0,$$
$$\alpha(0)=0. \tag{2.22}$$

Indeed, by (2.19) to (2.21), we get

$$\forall x\in V\quad |\alpha(x)|\leqq 2\varepsilon_4(x). \tag{2.23}$$

Furthermore, the function α satisfies, for all x in V,

$$2\alpha(x)\|x\|_H^2+\alpha(x)^2\langle T^{-1}x,x\rangle=2f(x)-\langle Tx,x\rangle, \tag{2.24}$$

so that we obtain,

$$\forall x\in V\quad f(x)=\tfrac{1}{2}\langle T\Phi(x),\Phi(x)\rangle, \tag{2.25}$$

where

$$\Phi(x)=x+\alpha(x)T^{-1}x. \tag{2.26}$$

(Let us point out that the mapping Φ is given by an explicit formula.)

Clearly the properties (2.6) and (2.7) are satisfied. We also have for $x\neq 0$ in V,

$$D\Phi(x)=Id+\alpha(x)T^{-1}+\langle D\alpha(x),\cdot\rangle T^{-1}x, \tag{2.27}$$

where, thanks to a differentiation of (2.24) with respect to x, $D\alpha(x)$ may be written as

$$D\alpha(x)=\frac{1}{\|x\|_H^2+\alpha(x)\langle T^{-1}x,x\rangle}(Df(x)-Tx-2\alpha(x)x-\alpha(x)^2T^{-1}x). \tag{2.28}$$

Therefore $D\alpha$ is in the space $C^0(V-\{0\};H)$ and $D\Phi$ obviously belongs to $C^0(V-\{0\};\mathscr{L}(H))$. Moreover, from the equality (2.27), we deduce that, for $x\neq 0$ in V,

$$\|D\Phi(x)-Id\|_{\mathscr{L}(H)}\leqq|\alpha(x)|\,\|T^{-1}\|_{\mathscr{L}(H)}+\|D\alpha(x)\|_H\,\|T^{-1}x\|_H. \tag{2.29}$$

Using (2.28) together with the hypothesis (H.3) and the property (2.23), we at once prove that, for $x\neq 0$ in V,

$$\|D\alpha(x)\|_H\,\|T^{-1}x\|_H\leqq C\|T^{-1}\|_{\mathscr{L}(H)}(\varepsilon_1(x)+\varepsilon_4(x)). \tag{2.30}$$

Finally from (2.29) and (2.30), we derive that $\|D\Phi(x)-Id\|_{\mathscr{L}(H)}$ tends to 0 as $\|x\|_X$ tends to 0, and the property (2.8) holds.

Now it remains to prove that Φ is a homeomorphism of a neighbourhood $\tilde{\mathscr{V}}$ of 0 in X onto a neighbourhood $\mathscr{W}$ of 0 in X. Let $\eta_0>0$ be a real number such that

$$\{x\in X;\ \|x\|_X<\eta_0\}\subset V, \tag{2.31}$$

and that we have, for $\|x\|_X<\eta_0$,

$$|\alpha(x)|\,\|T^{-1}\|_{\mathscr{L}(X)}<\tfrac{1}{2} \tag{2.32}$$

and

$$\|D\Phi(x)-Id\|_{\mathscr{L}(H)}<\tfrac{1}{2}. \tag{2.33}$$

We set

$$\tilde{\mathscr{V}}=\{x\in X;\ \|x\|_X<\eta_0\}\quad\text{and}\quad\mathscr{W}=\Phi(\tilde{\mathscr{V}}).$$

As

$$x-y=\Phi(x)-\Phi(y)-(\Phi(x)-\Phi(y)-(x-y)),$$

by (2.33), we obtain, for x and y in $\tilde{\mathcal{V}}$,

$$\|x-y\|_H \leqq 2\|\Phi(x)-\Phi(y)\|_H, \tag{2.34}$$

which proves that Φ is a one-to-one mapping from $\tilde{\mathcal{V}}$ onto $\mathcal{W}$. Let us now show that Φ^{-1} is a continuous mapping from $\mathcal{W}$ into $\tilde{\mathcal{V}}$. From the equation,

$$(\alpha(x)-\alpha(y))T^{-1}x=\Phi(x)-\Phi(y)-(Id_X+\alpha(y)T^{-1})(x-y),$$

we deduce, using the property (2.32) and the inequality (2.13) of Lemma 2.4, that

$$|\alpha(x)-\alpha(y)|\,\|T^{-1}x\|_H \leqq \|\Phi(x)-\Phi(y)\|_H+\tfrac{3}{2}\|x-y\|_H,$$

which becomes, owing to (2.34),

$$|\alpha(x)-\alpha(y)|\,\|T^{-1}x\|_H \leqq 4\|\Phi(x)-\Phi(y)\|_H. \tag{2.35}$$

But $x-y$ can be written as

$$x-y=\Phi(x)-\Phi(y)-\alpha(y)T^{-1}(x-y)-(\alpha(x)-\alpha(y))T^{-1}x,$$

so that, by (2.32),

$$\|x-y\|_X \leqq \|\Phi(x)-\Phi(y)\|_X+\tfrac{1}{2}\|x-y\|_X+|\alpha(x)-\alpha(y)|\,\|T^{-1}x\|_X. \tag{2.36}$$

From (2.35) and (2.36) we infer, for $x\neq 0$,

$$\|x-y\|_X \leqq 2\|\Phi(x)-\Phi(y)\|_X+8\frac{\|T^{-1}x\|_X}{\|T^{-1}x\|_H}\|\Phi(x)-\Phi(y)\|_X,$$

or

$$\|x-y\|_X \leqq C(x)\|\Phi(x)-\Phi(y)\|_X. \tag{2.37}$$

Inequality (2.37) proves that the mapping Φ^{-1} is continuous from X into X at the point $\Phi(x)$, for $x\neq 0$. But

$$y=\Phi(y)-\alpha(y)T^{-1}y,$$

so that, by using the property (2.32) again, we obtain

$$\|y\|_X \leqq 2\|\Phi(y)\|_X,$$

which proves that Φ^{-1} is continuous at the point 0.

Now we have shown that Φ is a homeomorphism of $\tilde{\mathcal{V}}$ onto $\mathcal{W}$; it remains to prove that $\mathcal{W}$ is a neighbourhood of 0 in X. To this end we are going to check that $\mathcal{W}\supset\frac{1}{2}\tilde{\mathcal{V}}$.

Let $y\neq 0$ be an element of X such that $\|y\|_X<\frac{1}{2}\eta_0$. We introduce the sequence $(x_n)_n$ defined by

$$\begin{aligned} &\text{(a)}\quad x_0=y,\\ &\text{(b)}\quad x_{n+1}=y-\alpha(x_n)T^{-1}x_n\equiv y-(\Phi(x_n)-x_n). \end{aligned} \tag{2.38}$$

Obviously x_n belongs to X, for all n; using the property (2.32), one also proves, by induction on n, that x_n belongs to $\tilde{\mathcal{V}}$, for all n. The relation (2.38)(b) implies, thanks to the property (2.33),

$$\|x_{n+1}-x_n\|_H<\tfrac{1}{2}\|x_n-x_{n-1}\|_H. \tag{2.39}$$

Therefore, we get, for all $n\in\mathbb{N}$,

$$\|x_{n+1}-x_n\|_H<\frac{1}{2^n}\|y\|_H. \tag{2.40}$$

From (2.40) it follows that $(x_n)_n$ is a Cauchy sequence in H. Hence, $(x_n)_n$ converges in the space H to an element $x \in H$. By (2.32), we have

$$\|y\|_H \leqq \|x_{n+1}\|_H + \tfrac{1}{2}\|x_n\|_H,$$

and also

$$\|x\|_H \geqq \tfrac{2}{3}\|y\|_H, \tag{2.41}$$

which means that $x \neq 0$. Furthermore, using the following equality

$$-(\alpha(x_{n+1}) - \alpha(x_n))T^{-1}x_n = (x_{n+2} - x_{n+1}) + \alpha(x_{n+1})T^{-1}(x_{n+1} - x_n), \tag{2.42}$$

as well as the properties (2.32) and (2.40), we obtain

$$|\alpha(x_{n+1}) - \alpha(x_n)|\,\|T^{-1}x_n\|_H \leqq \frac{1}{2^n}\|y\|_H.$$

Therefore, as x is not equal to zero, we get, thanks to Lemma 2.4,

$$|\alpha(x_{n+1}) - \alpha(x_n)| \leqq \frac{C(x)}{2^n}\|y\|_H, \tag{2.43}$$

where $C(x)$ is a positive constant (depending on x). From the equality

$$x_{n+1} - x_n = \alpha(x_n)T^{-1}(x_{n-1} - x_n) - (\alpha(x_n) - \alpha(x_{n-1}))T^{-1}x_{n-1}$$

we derive, thanks to the properties (2.32) and (2.43), that

$$\|x_{n+1} - x_n\|_X \leqq \frac{1}{2}\|x_n - x_{n-1}\|_X + \frac{C^*(x)}{2^n}, \tag{2.44}$$

where $C^*(x) > 0$ is a constant (depending on x). Therefore, we obtain, for all $n \in \mathbb{N}$,

$$\|x_{n+1} - x_n\|_X \leqq \frac{1}{2^n}\|x_1 - y\|_X + \frac{nC^*(x)}{2^n}, \tag{2.45}$$

which proves that $(x_n)_n$ is a Cauchy sequence in X. Therefore x belongs to $\tilde{V}$ and the relation (2.38)(b) gives us:

$$x = y - \alpha(x)T^{-1}x,$$

i.e.,

$$y = \Phi(x).$$

Proof of the Theorem 2.3. Part (2) in the proof of Theorem 2.1 may be followed up to (2.28); according to the implicit function theorem, we have only to prove that Φ belongs to $C^1(V; X)$ and even that $\|D\Phi(x) - Id_X\|_{\mathcal{L}(X)}$ tends to 0 as $\|x\|_X$ tends to 0.

Taking into account the equality (2.27), we get:

$$\|D\Phi(x) - Id\|_{\mathcal{L}(X)} \leqq |\alpha(x)|\,\|T^{-1}\|_{\mathcal{L}(X)} + \|D\alpha(x)\|_{X'}\|T^{-1}x\|_X. \tag{2.46}$$

Using (2.28) together with the hypothesis (H.3 bis), we at once prove that, for $x \neq 0$ in V,

$$\|D\alpha(x)\|_{X'} \leqq C(|\varepsilon_2(x)| + |\eta_2(x)|), \tag{2.47}$$

where $C > 0$ is a constant independent of x. From (2.46) and (2.47) it immediately follows that $\|D\Phi(x) - Id\|_{\mathcal{L}(X)}$ tends to 0 as $\|x\|_X$ tends to 0.

Remark 2.3. Let us recall that a mapping F from an open set $\mathcal{U}$ in X into H is "generator-differentiable" if, for any x in $\mathcal{U}$, there exists a continuous linear operator $DF(x)$ from X into H such that

$$\frac{\|F(x+h)-F(x)-DF(x)\cdot h\|_H}{\|h\|_H}\to 0 \quad \text{as } \|h\|_X\to 0. \tag{2.48}$$

F is said to be "C^1-generator-differentiable" if, moreover, $DF(x)$ belongs to $C^0(\mathcal{U};\mathcal{L}(X;H))$. J. Marsden has asked one of the authors if the homeomorphism Φ constructed in Theorem 2.1 is C^1-generator-differentiable. The answer is positive. As we have already proved that $D\Phi(x)$ belongs to $C^0(\mathcal{V};\mathcal{L}(H))$, it remains only to show that (2.48) holds for $F=\Phi$. But

$$\begin{aligned}\Phi(x+h)-\Phi(x)-D\Phi(x)h &= \alpha(x+h)T^{-1}h\\ &\quad+(\alpha(x+h)-\alpha(x)-\langle D\alpha(x),h\rangle)T^{-1}x,\end{aligned} \tag{2.49}$$

which implies, at the point $x\neq 0$, that

$$\begin{aligned}&\frac{\|\Phi(x+h)-\Phi(x)-D\Phi(x)h\|_H}{\|h\|_H}\\ &\leqq|\alpha(x+h)|\,\|T^{-1}\|_{\mathcal{L}(H)}+\frac{\left|\int_0^1\langle D\alpha(x+sh)-D\alpha(x),h\rangle\,ds\right|}{\|h\|_H}\end{aligned} \tag{2.50}$$

and also, because $D\alpha$ belongs to $C^0(V-\{0\},H)$, that

$$\begin{aligned}&\frac{\|\Phi(x+h)-\Phi(x)-D\Phi(x)h\|_H}{\|h\|_H}\\ &\leqq|\alpha(x+h)|\,\|T^{-1}\|_{\mathcal{L}(H)}+\sup_{0\leqq s\leqq 1}\|D\alpha(x+sh)-D\alpha(x)\|_H.\end{aligned} \tag{2.51}$$

From (2.51), we at once infer that (2.48) holds for $x\neq 0$. At the point $x=0$, (2.49) gives us:

$$\frac{1}{\|h\|_H}\|\Phi(h)-D\Phi(0)h\|_H\leqq|\alpha(h)|\,\|T^{-1}\|_{\mathcal{L}(H)},$$

and (2.48) still holds.

Remark 2.4. Assume that f and the inner product $\langle\cdot,\cdot\rangle$ are invariant with respect to the group representation $R:\Gamma\to GL(X)$, where Γ is a group. Then it is an easy matter to prove that the homeomorphism Φ is equivariant with respect to the representation R.

2.2. Let us now turn to parametrized versions of the second part of Theorem 2.1. Let $f(x,\lambda)$ be a functional defined on $X\times\Lambda$ where Λ is a Banach space. Assume that f satisfies the following hypotheses:

(B.1) There exist two neighbourhoods $\mathcal{V}_1$ of 0 in X and $\mathcal{V}_2$ of 0 in Λ such that $f\in C^1(\mathcal{V}_1\times\mathcal{V}_2;\mathbb{R})$ and $D_xf\in C^0(\mathcal{V}_1\times\mathcal{V}_2;H)$; furthermore $f(0,\lambda)=0$ and $D_xf(0,\lambda)=0$;

(B.2) There exists a continuous mapping $\lambda\to T_\lambda$ of $\mathcal{V}_2$ into $\mathcal{L}(H)$; moreover, $\forall\lambda\in\mathcal{V}_2$, T_λ is a symmetric operator (with respect to the inner product $\langle\cdot,\cdot\rangle$ of H) and, for $\lambda=0$, $T\equiv T_0$ is an isomorphism of X onto X;

(B.3) For all (x,λ) in $\mathcal{V}_1\times\mathcal{V}_2$, one has

$$\|D_xf(x,\lambda)-T_\lambda x\|_H\leqq\|x\|_H\varepsilon_1(x), \tag{2.52}$$

where

$$\lim_{\|x\|_X \to 0} \varepsilon_1(x) = 0.$$

PROPOSITION 2.5. *Assume that the hypotheses* (B.1), (B.2) and (B.3) *hold. Then there exists a homeomorphism* $\Psi(x, \lambda) = (\varphi(x, \lambda), \lambda)$ *of a neighbourhood* $\mathcal{V}$ *of* $(0, 0)$ *in* $X \times \Lambda$ *onto a neighbourhood* $\mathcal{W}$ *of* $(0, 0)$ *in* $X \times \Lambda$ *satisfying*

$$\forall (x, \lambda) \in \mathcal{V} \quad f(x, \lambda) = \tfrac{1}{2}\langle T_\lambda \varphi(x, \lambda), \varphi(x, \lambda)\rangle, \tag{2.53}$$

$$\varphi(0, \lambda) = 0, \qquad D_x\varphi(0, \lambda) = Id_H, \tag{2.54}$$

$$D_x\varphi \in C^0(\mathcal{V}; \mathcal{L}(H)). \tag{2.55}$$

Proof. Here we introduce the function $\alpha(x, \lambda)$ given by:

$$\alpha(0, \lambda) = 0 \quad \text{and, for } x \neq 0,$$

$$\alpha(x, \lambda) = \frac{2f(x, \lambda) - \langle T_\lambda x, x\rangle}{\sqrt{\langle T_\lambda T^{-1}x, x\rangle^2 + (2f(x, \lambda) - \langle T_\lambda x, x\rangle)\langle T_\lambda T^{-1}x, T^{-1}x\rangle} + \langle T_\lambda T^{-1}x, x\rangle} \tag{2.56}$$

and we set

$$\varphi(x, \lambda) = x + \alpha(x, \lambda) T^{-1}x. \tag{2.57}$$

Then, arguing as in the proof of Theorem 2.1, we show that $\Psi(x, \lambda)$ is a homeomorphism of $\mathcal{V}$ onto $\mathcal{W}$ and satisfies (2.53), (2.54) and (2.55). (The only difference is that here one has to check that $\Psi(x, \lambda)$ and $\Psi^{-1}(x, \lambda)$ are continuous mappings in λ; but this does not involve any difficulty.) □

Let $\tilde{\mathcal{V}}_1$ and $\tilde{\mathcal{V}}_2$ be two neighbourhoods of 0 in X and Λ, respectively, such that $\tilde{\mathcal{V}}_1 \times \tilde{\mathcal{V}}_2 \subset \mathcal{V}$; one can prove as in Remark 2.3 that, for $\lambda \in \tilde{\mathcal{V}}_2$, $\varphi(\cdot, \lambda)$: $x \in \tilde{\mathcal{V}}_1 \mapsto \varphi(x, \lambda)$ is C^1-generator differentiable.

From Proposition 2.5 one deduces at once the following result.

THEOREM 2.6. *Assume that the hypotheses* (B1), (B2) *and* (B3) *hold*; *assume furthermore that* $T_\lambda \in C^0(\mathcal{V}_2; \mathcal{L}(X))$. *Then there exists a homeomorphism* $\Psi^*(x, \lambda) = (\varphi^*(x, \lambda), \lambda)$ *of a neighbourhood* $\mathcal{V}^*$ *of* $(0, 0)$ *in* $X \times \Lambda$ *onto a neighbourhood* $\mathcal{W}^*$ *of* $(0, 0)$ *in* $X \times \Lambda$ *satisfying*:

$$\forall (x, \lambda) \in \mathcal{V}^* \quad f(x, \lambda) = \tfrac{1}{2}\langle T\varphi^*(x, \lambda), \varphi^*(x, \lambda)\rangle; \tag{2.58}$$

$$\varphi^*(0, \lambda) = 0, \qquad D_x\varphi^*(0, 0) = Id_X, \tag{2.59}$$

and

$$D_x\varphi^* \in C^0(\mathcal{V}^*; \mathcal{L}(H)). \tag{2.60}$$

Proof. The operator T_λ can be written as

$$T_\lambda = T(Id + T^{-1}(T_\lambda - T)).$$

There exists a neighbourhood $\mathcal{N}_2$ of 0 in Λ such that, for $\lambda \in \mathcal{N}_2$,

$$\|T^{-1}(T_\lambda - T)\|_{\mathcal{L}(X)} \leqq \tfrac{1}{2} \quad \text{and} \quad \|T^{-1}(T_\lambda - T)\|_{\mathcal{L}(H)} \leqq \tfrac{1}{2}.$$

Then we introduce the operator:

$$L_\lambda = Id + \sum_{k=1}^{\infty} c_k (T^{-1}(T_\lambda - T))^k, \tag{2.61}$$

where the real numbers c_k, $1 \leqq k \leqq \infty$, are defined by

$$(1 + x)^{1/2} = 1 + c_1 x + \cdots + c_k x^k + \cdots \quad \text{for } |x| < 1.$$

For λ in $\mathcal{N}_2$, L_λ is well defined by (2.61) and is an isomorphism of X onto X and of H onto H. Moreover, L_λ and $(L_\lambda)^{-1}$ belong to $C^0(\mathcal{N}_2; \mathcal{L}(X) \cap \mathcal{L}(H))$. One easily checks that

$$T_\lambda = L_\lambda^t T L_\lambda. \tag{2.62}$$

The mapping φ^* defined by

$$\varphi^*(x, \lambda) = L_\lambda \varphi(x, \lambda), \tag{2.63}$$

where φ is given by Proposition 2.5, clearly satisfies the requirements of Theorem 2.6. □

2.3. *Example.* We consider the functional

$$f(u, \lambda) = \int_\Omega a(\partial_0 u(x), \partial_1 u(x), \partial_2 u(x), x, \lambda)\, dx, \tag{2.64}$$

where Ω is a bounded open subset of $\mathbb{R}^2$ with a smooth enough boundary, $x = (x_1, x_2)$ is a generic point in $\mathbb{R}^2$, λ is a real parameter and where $\partial_0 u = u$, $\partial_i u = \partial u / \partial x_i$, for $i = 1, 2$. We assume that the function a belongs to the space $C^2(\mathbb{R}^6; \mathbb{R})$ and can be written as

$$a(u_0, u_1, u_2, x, \lambda) = \sum_{i,j=0}^{2} a_{ij}(u_0, u_1, u_2, x, \lambda) u_i u_j$$

where $a_{ij} \in C^0(\mathbb{R}^6; \mathbb{R})$ and $a_{ij} = a_{ji}$, for $i \neq j$. The above equality means that the function a and its first partial derivatives (with respect to u_0, u_1 and u_2) vanish at the points $(0, 0, 0, x, \lambda) \equiv (0, x, \lambda)$. We suppose that there exists a real number $\alpha > 0$ such that

$$\forall \xi \in \mathbb{R}^2 \quad \sum_{i,j=0}^{2} a_{ij}(0, x, 0) \xi_i \xi_j \geqq 2\alpha \sum_{i=1}^{2} \xi_i^2 \tag{2.65}$$

and that

$$\left.\begin{array}{l} u \in H_0^1(\Omega) \quad \text{and} \quad \forall v \in H_0^1(\Omega) \\ \displaystyle\int_\Omega \sum_{i,j=0}^{2} a_{ij}(0, x, 0) \partial_i u \partial_j v\, dx = 0 \end{array}\right\} \Rightarrow u = 0. \tag{2.66}$$

We set $X = H_0^1(\Omega) \cap W^{1,\infty}(\Omega)$, $H = H_0^1(\Omega)$, $\Lambda = \mathbb{R}$. We now introduce the following inner product on H:

$$\langle u, v \rangle = 2 \int_\Omega \sum_{i,j=0}^{2} a_{ij}(0, x, 0) \partial_i u \partial_j v\, dx + \beta \int_\Omega uv\, dx, \tag{2.67}$$

where $\beta > 0$ is a real number such that, for all v in $H_0^1(\Omega)$,

$$\langle v, v \rangle \geqq \alpha \int_\Omega \sum_{i=0}^{2} (\partial_i v)^2\, dx.$$

The assumption (2.65) insures the existence of such a number β. We are now able to formulate our last assumption; we suppose that the operator $G \in \mathcal{L}(H)$ given by

$$\left.\begin{array}{l} Gu \in H_0^1(\Omega) \quad \text{and} \quad \forall v \in H_0^1(\Omega), \\ \langle Gu, v \rangle = \displaystyle\int_\Omega uv\, dx, \end{array}\right. \tag{2.68}$$

is a compact operator of $\mathcal{L}(X)$. (This property is true in particular if the coefficients $a_{ij}(0, x, 0)$ belong to the Hölder space $C^\delta(\Omega)$ for $0 < \delta \leqq 1$.) Let us verify that the above

functional f satisfies the hypotheses (B.1), (B.2) and (B.3). It is clear that f belongs to $C^2(X\times\Lambda;\mathbb{R})$. We introduce the linear operators T_λ defined by

$$T_\lambda u\in H_0^1(\Omega) \quad \text{and, for all } v \text{ in } H_0^1(\Omega),$$

$$\langle T_\lambda u, v\rangle = 2\int_\Omega \sum_{i,j=0}^{2} a_{ij}(0,x,\lambda)\partial_i u\partial_j v\,dx.$$

It is not difficult to prove that the property (B.1) is satisfied, T_λ belongs to $C^0(\mathbb{R};\mathscr{L}(H))$ and (B.3) holds. In order to prove (B.2), it remains to prove that $T\equiv T_0\in\mathscr{L}(X)$ and is an isomorphism of X onto X. We remark that $T=Id-\beta G$. As G is a compact operator of $\mathscr{L}(X)$, the assumption (2.66) implies that T is an isomorphism of X onto X. Thus the hypotheses (B.1), (B.2) and (B.3) are satisfied and the Proposition 2.5 applies.

3. The splitting lemma.

3.1. Now we give the splitting lemma which allows us to replace infinite-dimensional problems by finite-dimensional ones. In particular, it allows us to reduce the study of the minima of a C^2 functional defined on a infinite-dimensional Banach space to the study of the minima of a functional defined on a finite-dimensional space. But, contrary to Magnus [9] and to Golubitsky and Marsden [5], we need not use the Morse lemma in our proofs. However, at the end of the section we shall show how to use our Morse lemma (Theorem 2.6) in the proof.

We keep the notation of § 2. Let $f(x,\lambda)$ be a functional defined on $X\times\Lambda$ where Λ is a Banach space. Assume that the following properties hold:

(h.1) $\quad f\in C^p(X\times\Lambda;\mathbb{R})$ and $D_x f\in C^p(X\times\Lambda;X)$ with $p\geqq 1$,

(h.2) $\quad D_x f(0,0)=0$,

and

(h.3) there exist two closed subspaces Z and V of X satisfying

(3.1) $\quad X=Z\oplus V$, where V and Z are orthogonal with respect to the inner product $\langle\cdot,\cdot\rangle$,

and the restriction $\tilde{T}$ of $T\equiv D_{xx}^2 f(0,0)$ to Z is an isomorphism of Z onto Z.

Remark 3.1. The condition (3.1) holds, for instance, if T is a Fredholm operator from X into X of index 0 (with $V=\operatorname{Ker} T$ and $Z=\operatorname{Im} T$).

Let F be the C^p mapping from $Z\times V\times\Lambda$ into Z defined by

$$F(z;v,\lambda)=P_Z D_x f(z+v,\lambda), \tag{3.2}$$

where P_Z denotes the linear projector from X onto Z corresponding to the decomposition $X=Z\oplus V$. We remark that $F(0;0,0)=0$ and that $D_z F(0;0,0)=\tilde{T}$ is an isomorphism of Z onto Z; therefore, by applying the implicit function theorem, we obtain the following result.

LEMMA 3.1. *There exist three neighbourhoods $\mathscr{Z}$, $\mathscr{V}$ and $\mathscr{L}$ of 0 in Z, V and Λ, respectively, and, for any $(v,\lambda)\in\mathscr{V}\times\mathscr{L}$, a unique element $H(v,\lambda)\in\mathscr{Z}$ such that*

$$P_Z D_x f(H(v,\lambda)+v,\lambda)=0. \tag{3.3}$$

Moreover the mapping H belongs to $C^p(\mathscr{V}\times\mathscr{L};Z)$ and satisfies

$$H(0,0)=0, \qquad D_v H(0,0)=0. \tag{3.4}$$

Now, if $V\neq\{0\}$, we introduce the following "splitting"; we set, for $(v,\lambda)\in\mathscr{V}\times\mathscr{L}$,

$$\begin{aligned} f_V(v,\lambda) &= f(H(v,\lambda)+v,\lambda), \\ f_Z(z;v,\lambda) &= f(z+H(v,\lambda)+v,\lambda)-f_V(v,\lambda). \end{aligned} \tag{3.5}$$

Of course,

$$f(z+H(v,\lambda)+v,\lambda)=f_Z(z;v,\lambda)+f_V(v,\lambda). \tag{3.6}$$

Remark 3.2. For any $(v,\lambda)\in\mathcal{V}\times\mathcal{L}$, we have

$$f_Z(0;v,\lambda)=0; \tag{3.7}$$

and, as V and Z are orthogonal with respect to $\langle\cdot,\cdot\rangle$, (3.3) implies, for any $(v,\lambda)\in\mathcal{V}\times\mathcal{L}$,

$$\forall z^*\in Z,\quad D_z f_Z(0;v,\lambda)z^*=0, \tag{3.8}$$

and

$$\forall v^*\in V\quad D_v f_V(v,\lambda)w=\langle D_x f(H(v,\lambda)+v,\lambda),w\rangle. \tag{3.9}$$

Moreover, f_Z and $D_z f_Z$ belong to $C^p(\mathcal{X}\times\mathcal{V}\times\mathcal{L};\mathbb{R})$ and $C^p(\mathcal{X}\times\mathcal{V}\times\mathcal{L};X)$, respectively; and f_V and $D_v f_V$ belong to $C^p(\mathcal{V}\times\mathcal{L};\mathbb{R})$ and $C^p(\mathcal{V}\times\mathcal{L};X)$, respectively.

THEOREM 3.2. *For any $\lambda\in\mathcal{L}$, the mapping $v\to H(v,\lambda)+v$ is a bijection between the critical points of $f_V(\cdot,\lambda)$ in $\mathcal{V}$ and those of $f(\cdot,\lambda)$ in $\mathcal{X}\oplus\mathcal{V}$.*

Proof. It is an obvious consequence of Lemma 3.1 and of relation (3.9).

Now we introduce the additional hypothesis:

(h.4) The operator T is positive on Z, i.e.,

$$\forall z\in Z\quad \langle Tz,z\rangle\geqq 0. \tag{3.10}$$

LEMMA 3.3. *Assume that the hypotheses* (h.1), (h.2), (h.3) *and* (h.4) *hold. Then we can choose the neighbourhoods $\mathcal{X}$, $\mathcal{V}$, $\mathcal{L}$ in Lemma* 3.1 *in such a way that*

$$\forall(z,v,\lambda)\in\mathcal{X}\times\mathcal{V}\times\mathcal{L}\quad \text{with } z\neq 0,\quad f_Z(z;v,\lambda)>0. \tag{3.11}$$

Proof. Using a Taylor formula, we show, thanks to (3.7) and (3.8), that there exists a real number $t\in]0,1[$, depending on z, v and λ, such that

$$f_Z(z;v,\lambda)=\tfrac{1}{2}\langle D^2_{zz}f_Z(tz;v,\lambda)z,z\rangle$$

or also

$$f_Z(z;v,\lambda)=\tfrac{1}{2}\langle Tz,z\rangle+\tfrac{1}{2}\langle B_t z,z\rangle, \tag{3.12}$$

where

$$B_t=D^2_{xx}f(tz+H(v,\lambda)+v,\lambda)-T. \tag{3.13}$$

The Cauchy-Schwarz inequality

$$\langle Ty,z\rangle^2\leqq\langle Ty,y\rangle\langle Tz,z\rangle,$$

with $y=\tilde{T}^{-1}z$, $z\in Z$, implies

$$\langle Tz,z\rangle\langle\tilde{T}^{-1}z,z\rangle\geqq\|z\|_H^4;$$

therefore, thanks to Lemma 2.4, we obtain

$$\forall z\in Z\quad \langle Tz,z\rangle\geqq\frac{\|z\|_H^2}{\|\tilde{T}^{-1}\|_{\mathcal{L}(Z)}}. \tag{3.14}$$

On the other hand, using Lemma 2.4 once more, we get

$$|\langle B_t z, z\rangle| \leqq \|B_t\|_{\mathscr{L}(X)} \|z\|_H^2. \tag{3.15}$$

Now we at once deduce the property (3.11) from (3.14) and (3.15) by choosing the neighbourhoods $\mathscr{Z}$, $\mathscr{V}$ and $\mathscr{L}$, introduced in Lemma 3.1, in such a way that

$$\forall t \in [0,1] \quad \forall (z, v, \lambda) \in \mathscr{Z} \times \mathscr{V} \times \mathscr{L} \quad \|B_t\|_{\mathscr{L}(X)} < \frac{1}{\|\tilde{T}^{-1}\|_{\mathscr{L}(Z)}}. \tag{3.16}$$

The following theorem is an obvious consequence of Theorem 3.2 and the previous lemma.

THEOREM 3.4. *Assume that the hypotheses* (h.1), (h.2), (h.3) *and* (h.4) *hold and that the neighbourhoods $\mathscr{Z}$, $\mathscr{V}$, $\mathscr{L}$ are chosen in such a way that* (3.11) *is satisfied. Then, for any $\lambda \in \mathscr{L}$, the mapping $v \to H(v, \lambda) + v$ is a bijection between the minima of $f_V(\cdot, \cdot)$ in $\mathscr{V}$ and those of $f(\cdot, \lambda)$ in $\mathscr{Z} \oplus \mathscr{V}$.*

3.2. A few comments. (1) We can also prove Lemma 3.3 and Theorem 3.4 by using the Morse lemma given in § 2. Indeed, applying the Theorem 2.6 to the function f_Z, we obtain the following result.

THEOREM 3.5. *Assume that the hypotheses* (h.1), (h.2), (h.3) *hold. Then there exists a homeomorphism $\Psi^*(z, v, \lambda) = (\varphi^*(z, v, \lambda), v, \lambda)$ of a neighourhood $\mathscr{Z}^* \times \mathscr{V}^* \times \mathscr{L}^*$ of 0 in $Z \times V \times \Lambda$ onto another neighbourhood of 0 in $Z \times V \times \Lambda$ satisfying*

$$\forall (z, v, \lambda) \in \mathscr{Z}^* \times \mathscr{V}^* \times \mathscr{L}^* \quad f_Z(z, v, \lambda) = \tfrac{1}{2}\langle T\varphi^*(z, v, \lambda), \varphi^*(z, v, \lambda)\rangle \tag{3.17}$$

and

$$\varphi^*(0, v, \lambda) = 0, \qquad D_z\varphi^*(0, v, \lambda) = Id_Z. \tag{3.18}$$

(Let us remark that Theorem 3.5 is a generalization of the "generalized Morse lemma" of Mawhin and Willem [10].)

Now Lemma 3.3 is a direct consequence of Theorem 3.5 and of the hypothesis (h.4).

Of course this second proof of Lemma 3.3 is shorter. Nevertheless the first method of proof admits more generalizations than the second one and is well adapted to approximate problems encountered in numerical analysis (see [11, Chap. II] for such a generalization). In particular, the first method of proof enables us to show Theorem 3.4 when $\nabla f(0, 0)$ is near zero, but does not vanish (see [11, Chap. II]).

(2) When $V = \operatorname{Ker} T$ and $Z = \operatorname{Im} T$, the Theorem 3.2 is the variational counterpart of the Lyapunov-Schmidt procedure on the Euler equation $D_x f(x, \lambda) = 0$ and, of course, the two procedures give the same change of coordinates. But our point of view also gives a bijection between the minima of $f_V(\cdot, \lambda)$ and those of $f(\cdot, \lambda)$. In [2], a simplified procedure for the computation of the Taylor expansion of $f_V(\cdot, \lambda)$ is given: it uses the Faa di Bruno formula.

(3) Examples of applications of the splitting lemma can be found in [1], [2], [3] and [9]. The splitting lemma is especially useful in the problems of elasticity. In [11, Chap. III], one applies the splitting lemma and its discrete version to the elasticity problem studied in [3] and to its approximation, respectively.

REFERENCES

[1] M. BUCHNER, J. MARSDEN AND S. SCHECTER, *Examples for the infinite dimensional Morse lemma*, this Journal, 14 (1983), pp. 1045-1055.

[2] E. BUZANO AND G. GEYMONAT, *Geometrical methods in some bifurcation problems of elasticity*, in Fifth Symposium on Trends of Applications of Pure Mathematics to Mechanics, Springer Lecture Notes in Physics 195, Springer-Verlag, Berlin, New York, Heidelberg, 1984, pp. 5-19.

[3] E. BUZANO, G. GEYMONAT AND T. POSTON, *Post buckling behaviour of a non-linearly hyperelastic thin rod with cross-section invariant under the dihedral group D_n*, Arch. Rational Mech. Anal., 89 (1985), pp. 307–388.

[4] J. DIEUDONNE, *Eléments d'Analyse*, Vol. I, Gauthier-Villars, 1969.

[5] M. GOLUBITSKY AND J. MARSDEN, *The Morse lemma in infinite dimensions via singularity theory*, this Journal, 14 (1983), pp. 1037–1044.

[6] T. HUGHES AND J. MARSDEN, *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[7] N. H. KUIPER, *C^1-equivalence of functions near isolated critical points*, in Symposium on Infinite-Dimensional Topology, Baton Rouge, LA, 1967.

[8] J. L. LIONS AND J. PEETRE, *Sur une classe d'espaces d'interpolation*, Publ. Math. I.H.E.S., 19 (1964), pp. 5–68.

[9] R. J. MAGNUS, *A splitting lemma for nonreflexive Banach spaces*, Math. Scan., 46 (1980), pp. 118–128.

[10] J. MAWHIN AND M. WILLEM, *On the generalized Morse lemma*, Séminaire de Mathématique (nouvelle série), 1er semestre 1985, Institut de Mathématique Pure et Appliquée, Université Catholique de Louvain.

[11] G. RAUGEL, *Approximation numérique de problèmes non linéaires*, Thèse d'Etat, Université de Rennes, France, 1984.

[12] A. J. TROMBA, *Almost Riemannian structures on Banach manifolds, the Morse lemma and the Darboux lemma*, Canad. J. Math., 28 (1976), pp. 640–652.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

009

BOUNDS FOR THE TAILS OF SHARP-CUTOFF FILTER KERNELS*

B. F. LOGAN†

Abstract. In communication theory, it is convenient to deal with bandlimited signals obtained by convolving an arbitrary bounded function with a filter kernel $k(t; \alpha, \beta)$ whose Fourier transform is 1 over the interval $(-\alpha, \alpha)$, and vanishes outside the interval $(-\beta, \beta)$, $0 < \alpha < \beta < \infty$. For α/β near 1, the sharp-cutoff case, the L_1-norm of the kernel must be large. In this paper, estimates are given for $\int_{|t|>T} |k(t; \alpha, \beta)|\, dt$, the norm in the tails of the kernel, which show that T must grow like $(\beta - \alpha)^{-1}$ as $\alpha \to \beta$ in order for the norm in the tails to be (say) less than 1. This result confirms a conjecture of J. C. Lagarias and A. M. Odlyzko who used such filter kernels in a method for computing $\pi(x)$, the number of primes not exceeding x.

Key words. bandlimiting, Lebesgue constants, L_1-norm

AMS(MOS) subject classification. 42A05

The notion of bandlimiting time signals by ideal filtering is pervasive in communication theory, owing to the fact that the resulting signal can be represented by a countable number of data or "samples". For a suitable class of signals, e.g., L_2, this bandlimiting can be accomplished (in theory) by convolving the signal with the filter kernel

$$k(t) = \frac{\sin \alpha t}{\pi t}. \tag{1}$$

However, this kernel is not L_1 (absolutely integrable) so the convolution with arbitrary bounded signals is not always defined. In order for a kernel to belong to L_1, it is necessary that its Fourier transform be continuous and tend to zero at $\pm\infty$. Here we consider filters characterized by kernels $k(t; \alpha, \beta)$ in L_1 whose Fourier transforms $\hat{k}$ satisfy

$$\hat{k}(\omega; \alpha, \beta) = \int_{-\infty}^{\infty} k(t; \alpha, \beta)\, e^{-i\omega t}\, dt = \begin{cases} 1 & \text{for } -\alpha \leqq \omega \leqq \alpha, \\ 0 & \text{for } |\omega| \geqq \beta \end{cases} \tag{2}$$

where $0 < \alpha < \beta < \infty$. The collection of such kernels will be denoted by $K(\alpha, \beta)$. If $(\beta - \alpha)$, the "cut-off interval," is small compared to $(\beta + \alpha)$ or, say α, then it is quite easy to show that the L_1 norm of $k(t; \alpha, \beta)$ must be large. J. C. Lagarias and A. M. Odlyzko [1], in using sharp cutoff filters in a method to compute $\pi(x)$, the number of primes not exceeding x, raised the following question concerning the norm in the tails of the kernels. If $\alpha \to \beta$, how small can one take T (depending on α and β) such that (for the best choice of $k(t; \alpha, \beta)$)

$$\int_{|t|>T} |k(t; \alpha, \beta)|\, dt \leqq 1. \tag{3}$$

They conjectured that $(\beta - \alpha)T$ could not tend to zero with (3) holding for any k in $K(\alpha, \beta)$. We show here that the conjecture is true. One would really like to determine

$$m_T(\alpha, \beta) = \inf_{k \in K(\alpha,\beta)} \int_{|t|>T} |k(t; \alpha, \beta)|\, dt. \tag{4}$$

This appears to be a very difficult problem to solve (except in special cases with $T = 0$), but inequalities are fairly simple to obtain which answer the question at hand.

* Received by the editors October 14, 1986; accepted for publication April 12, 1987.
† AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

It is convenient to rescale the kernels so that they may be described in terms of one parameter λ. We denote by $K_\lambda = K(\lambda-1, \lambda+1)$ the collection of kernels $k(t)$ in L_1 whose Fourier transforms $\hat{k}$ satisfy

$$\hat{k}(\omega) = \int_{-\infty}^{\infty} k(t)\, e^{-i\omega t}\, dt = \begin{cases} 1 & \text{for } -(\lambda-1) \leqq \omega \leqq \lambda - 1, \\ 0 & \text{for } |\omega| \geqq \lambda + 1 \end{cases} \tag{5}$$

where $\lambda > 1$. The "cut-off interval" is now fixed at 2, and we are interested in the case where $\lambda \to \infty$. We wish to obtain inequalities for the quantity defined by

$$\mu(\lambda, c) = \inf_{k \in K_\lambda} \int_{|t|>c} |k(t)|\, dt. \tag{6}$$

Now any kernel in L_1 of the form

$$k(t) = g(t)\frac{\sin \lambda t}{\pi t}, \qquad \lambda > 1 \tag{7}$$

where the Fourier transform of g vanishes outside $[-1, 1]$ and $g(0) = 1$, belongs to K_λ. This is not the most general representation for kernels in k_λ, but it suffices for obtaining an upper bound for $\mu(\lambda, c)$. Let $B(1)$ denote the collection of functions g which are restrictions to the real line of entire functions of exponential type 1, which, in engineering terminology, are the bandlimited functions whose (generalized) Fourier transforms vanish outside $[-1, 1]$. Then from (6) and (7) we have

$$\mu(\lambda, c) \leqq \inf_{\substack{g \in B(1) \\ g(0)=1}} \frac{1}{\pi} \int_{|t|>c} \frac{|g(t)|}{|t|} \cdot |\sin \lambda t|\, dt. \tag{8}$$

It has been shown elsewhere [2] that

$$\inf_{\substack{g \in B(1) \\ g(0)=1}} \frac{1}{\pi} \int_{|t|>c} \frac{|g(t)|}{|t|}\, dt = \frac{2}{\pi} \log \frac{1+e^{-c}}{1-e^{-c}}, \tag{9}$$

the extremal function in (9) being

$$g(t) = g(t; c) = \frac{c}{\sinh c} \frac{\sin \sqrt{t^2 - c^2}}{\sqrt{t^2 - c^2}}. \tag{10}$$

Thus we have

$$\mu(\lambda, c) < \frac{2}{\pi} \log \frac{1+e^{-c}}{1-e^{-c}}. \tag{11}$$

In (8) we can expand $|\sin \lambda t|$ in a Fourier series (absolutely convergent)

$$|\sin \lambda t| = \frac{2}{\pi} - a_1 \cos 2\lambda t - a_2 \cos 4\lambda t - \cdots, \tag{12}$$

and then as $\lambda \to \infty$ only the constant term $2/\pi$ will contribute to the integral, provided $c > 0$ is fixed. We have equality in (9) for $g(t) = g(t; c)$; hence

$$\mu(\lambda, c) \leqq \frac{1}{\pi} \int_{|t|>c} \frac{|g(t; c)|}{|t|} \cdot |\sin \lambda t|\, dt \sim \frac{4}{\pi^2} \log \frac{1+e^{-c}}{1-e^{-c}} \qquad (\lambda \to \infty), \quad (c > 0). \tag{13}$$

It is believed that this is the correct asymptotic behavior of $\mu(\lambda; c)$ as $\lambda \to \infty$, with c held fixed, $c > 0$, but we will not attempt to prove that here. (It would be true if the extremal function had the simple form (7) for all c and λ, or tended in norm to that form as $\lambda \to \infty$.)

Now lower bounds for $\mu(\lambda, c)$ are more difficult to obtain. What we need here are functions $s(t; \lambda, c)$ of unit norm in L_∞ which vanish over $(-c, c)$ and have spectral gaps $(-\lambda_2, -\lambda_1)$ and (λ_1, λ_2), where

$$\lambda_1 = \lambda - 1, \quad \lambda_2 = \lambda + 1 \quad (\lambda > 1).$$

Such functions $s(t; \lambda, c)$ have the representation

$$s(t; \lambda, c) = f(t) + h(t), \tag{14}$$

where $s(t; \lambda, c) = 0$, $|t| \leqq c$, $\sup_t |s(t; \lambda, c)| = 1$, and $f(t)$ is a bounded bandlimited function whose "Fourier transform" vanishes outside $[-\lambda_1, \lambda_1]$, and $h(t)$ is a bounded high-pass function whose "Fourier transform" vanishes over $(-\lambda_2, \lambda_2)$. Since the Fourier transform of any kernel k in K_λ is 1 over the interval $[-\lambda_1, \lambda_1]$ and vanishes outside $(-\lambda_2, \lambda_2)$, the convolution of $s(t; \lambda, c)$ in (14) with $k(t)$ [or $k(-t)$] simply gives $f(t)$, rejecting $h(t)$. In particular,

$$\int_{-\infty}^{\infty} k(t)s(t; \lambda, c)\, dt = \int_{|t| \geqq c} k(t)s(t; \lambda, c)\, dt = f(0), \qquad k \in K_\lambda \tag{15}$$

gives the inequality

$$|f(0)| \leqq \int_{|t| > c} |k(t)|\, dt, \qquad k \in K_\lambda. \tag{16}$$

Thus we would like to find $s(t; \lambda, c)$ of the form (14), where $f(0)$ is as large as possible. This presents another difficult problem except in the case $c = 0$, $\lambda = n$, where n is an integer not less than 2. In this case the optimal function is

$$s(t; n, 0) = \operatorname{sgn}\left\{\frac{\sin nt}{\sin t}\right\}, \qquad n \geqq 2 \tag{17}$$

and the kernel of minimal norm in K_n is

$$k_n(t; 0) = \frac{\sin t}{t}\frac{\sin nt}{\pi t}. \tag{18}$$

To see this, we first note that

$$s(t + \pi; n, 0) = (-1)^{n-1} s(t; n, 0). \tag{19}$$

So in the case where n is odd ($\geqq 3$), $s(t; n, 0)$ has period π and a Fourier series of the form

$$s(t; n, 0) = b_0(n) + 2\sum_{k=1}^{\infty} b_{2k}(n) \cos 2kt, \quad n \text{ odd} \geqq 3. \tag{20}$$

In the case where n is even ($\geqq 2$), $s(t; n, 0)$ has period 2π but, in accord with (19), has only odd harmonics

$$s(t; n, 0) = 2\sum_{k=0}^{\infty} b_{2k+1}(n) \cos (2k+1)t, \quad n \text{ even} \geqq 2. \tag{21}$$

In either case, $s(t; n, 0)$ has the spectral gaps $(n-1, n+1)$ and $(-n-1, -n+1)$. So if k is any kernel in K_n we have

$$\begin{aligned} \int_{-\infty}^{\infty} s(t; n, 0)k(t)\, dt &= \int_{-\infty}^{\infty} s(t; n, 0)k_n(t; 0)\, dt \\ &= \int_{-\infty}^{\infty} |k_n(t; 0)|\, dt = \int_{-\infty}^{\infty} \frac{\sin^2 t}{\pi t^2}\left|\frac{\sin nt}{\sin t}\right| dt \leqq \int_{-\infty}^{\infty} |k(t)|\, dt. \end{aligned} \tag{22}$$

Now note in the next-to-last integral in (22), on expanding the periodic function $|\cdot|$ (period π) in a Fourier series, that only the constant term contributes to the integral. Thus we obtain the result

$$\mu(n,0)=\frac{2}{\pi}\int_0^{\pi/2}\left|\frac{\sin nt}{\sin t}\right|dt, \qquad n\geqq 1$$
$$=\frac{4}{\pi^2}\log n+O(1), \qquad n\to\infty. \tag{23}$$

(See "Lebesgue Constants," [3, p. 67].)

Now if $k(t)$ belongs to $K_\lambda = K(\lambda-1,\lambda+1)$, then $ak(at)$ belongs to $K(a(\lambda-1), a(\lambda+1))$. So setting

$$a=\frac{\alpha}{\lambda-1}, \quad \frac{\lambda+1}{\lambda-1}=\frac{\beta}{\alpha} \quad \text{or} \quad \lambda=\frac{\beta+\alpha}{\beta-\alpha},$$

we see that the connection between $m_T(\alpha,\beta)$ defined in (4) and $\mu(\lambda, c)$ defined in (6) is

$$m_T(\alpha,\beta)=\mu\left(\frac{\beta+\alpha}{\beta-\alpha},(\beta-\alpha)\frac{T}{2}\right). \tag{24}$$

Now we wish to show that $m_T(\alpha,\beta)\to\infty$ as $\alpha\to\beta$ if $(\beta-\alpha)T\to 0$; i.e., we need a lower bound for $\mu(\lambda, c)$ for small c and large λ.

For fixed c, it would seem that $\mu(\lambda, c)$ should be an increasing function of λ. We are not able to show that here, but we only need the "quasi-monotonicity" (29) to obtain the desired result from lower bounds for $\mu(n, c)$, $c<\pi/2$.

First suppose $n\leqq\lambda<n+1$ (n an integer $\geqq 2$). Then if $k_\lambda(t)$ belongs to $K(\lambda-1, \lambda+1)$, the dilation

$$k_n(t)=\frac{n+1}{\lambda+1}k_\lambda\left(\frac{n+1}{\lambda+1}t\right) \tag{25}$$

belongs to $K((n+1)(\lambda-1)/(\lambda+1), n+1)\subseteq K(n-1, n+1)=K_n$. We have

$$\inf_{k_\lambda\in K_\lambda}\int_{|t|>c}|k_\lambda(t)|\,dt=\mu(\lambda, c). \tag{26}$$

Using (25) we have

$$\int_{|t|>c}|k_n(t)|\,dt=\int_{|t|>c}\frac{n+1}{\lambda+1}\left|k_\lambda\left(\frac{n+1}{\lambda+1}t\right)\right|dt\geqq\mu(n, c) \qquad (n\leqq\lambda<n+1). \tag{27}$$

Thus

$$\int_{|t|>(n+1)c/(\lambda+1)}|k_\lambda(t)|\,dt\geqq\mu(n, c);$$

whence follows

$$\mu\left(\lambda,\frac{n+1}{\lambda+1}c\right)\geqq\mu(n, c), \qquad n\leqq\lambda<n+1, \tag{28}$$

or

$$\mu(\lambda, c)\geqq\mu\left(n,\frac{\lambda+1}{n+1}c\right), \qquad n\leqq\lambda<n+1. \tag{29}$$

It is sufficient then to show that

$$\mu(n, c) > \frac{2}{\pi}\int_c^{\pi/2} \left|\frac{\sin nt}{\sin t}\right| dt, \qquad n \geqq 2, \quad 0 < c < \frac{\pi}{2}. \tag{30}$$

To obtain the last result, define the periodic function

$$\begin{aligned} \sigma_n(t; c) &= \operatorname{sgn}\left\{\frac{\sin nt}{\sin t}\right\}, \qquad c \leqq t \leqq \pi - c \\ &= 0, \qquad -c < t < c, \end{aligned} \tag{31}$$

$$\sigma_n(t+\pi; c) = (-1)^{n-1}\sigma_n(t; c), \qquad -\infty < t < \infty. \tag{31a}$$

That is, we obtain $\sigma_n(t; c)$ by subtracting from $s(t; n, 0)$, defined in (17), its restriction to the intervals $(-c, c)$ modulo π. For n either odd or even, $\sigma_n(t; c)$ has the same form of Fourier series as $s_n(t; n, 0)$. In particular $\sigma_n(t; c)$ has spectral gaps $(n-1, n+1)$ and $(-n-1, -n+1)$. Hence for any k in K_n, we have

$$\int_{-\infty}^{\infty} k(t)\sigma_n(t; c)\, dt = \int_{-\infty}^{\infty} \frac{\sin nt}{\pi t}\frac{\sin t}{t}\sigma_n(t; c)\, dt. \tag{32}$$

So, since $|\sigma_n(t; c)| \leqq 1$ for $|t| > c$, vanishing over $(-c, c)$ and translates of this interval, we have for $c > 0$

$$\begin{aligned} \int_{|t|>c} |k(t)|\, dt &> \left|\int_{-\infty}^{\infty} k(t)\sigma_n(t; c)\, dt\right| \\ &= \int_{|t|>c} \frac{\sin nt}{\pi t}\frac{\sin t}{t}\sigma_n(t; c)\, dt \\ &= \int_{|t|>c} \frac{\sin^2 t}{\pi t^2}\frac{\sin nt}{\sin t}\sigma_n(t; c)\, dt \\ &= \frac{2}{\pi}\int_c^{\pi/2} \left|\frac{\sin nt}{\sin t}\right| dt \qquad \left(k \in K_n, 0 < c < \frac{\pi}{2}\right). \end{aligned} \tag{33}$$

Thus

$$\mu(n, c) > \frac{2}{\pi}\int_c^{\pi/2} \left|\frac{\sin nt}{\sin t}\right| dt \sim \frac{4}{\pi^2}\log\left(\cot\frac{c}{2}\right), \qquad n \to \infty, \quad 0 < c < \frac{\pi}{2}. \tag{34}$$

Now if $0 < c < \varepsilon$ for any positive $\varepsilon < \pi/2$, we have

$$\mu(\lambda, c) > \mu(\lambda, \varepsilon). \tag{35}$$

Recalling the relation (24) between $m_T(\alpha, \beta)$ and $\mu(\lambda, c)$, and the inequalities (29) and (34), we see that if T is "little-oh" of $(\beta - \alpha)^{-1}$ as $\alpha \to \beta$, then $m_T(\alpha, \beta) \to \infty$.

REFERENCES

[1] J. C. LAGARIAS AND A. M. ODLYZKO, *Computing $\pi(x)$: An analytic method*, J. Algorithms, 8 (1987), pp. 173-191.

[2] B. F. LOGAN, *Optimal truncation of the Hilbert transform kernel for bounded high-pass functions*, in Proc. Fifth Annual Princeton Conference on Information Sciences and Systems, Dept. of Elec. Engrg., Princeton University, Princeton, NJ, 1971, pp. 10-12.

[3] A. ZYGMUND, *Trigonometric Series*, Vol. I, 2nd ed., Cambridge University Press, London, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

010

SOME ISOPERIMETRIC INEQUALITIES FOR THE LEVEL CURVES OF CAPACITY AND GREEN'S FUNCTIONS ON CONVEX PLANE DOMAINS*

MARCO LONGINETTI†

Abstract. The perimeter and the area of the convex level sets of capacity and Green's functions in convex plane domains are shown to satisfy sharp differential inequalities. Isoperimetric inequalities for capacity problems for optimal conductors are derived.

Key words. level set analysis, isoperimetric inequalities, capacity and Green's function, convexity

AMS(MOS) subject classifications. 31A15, 31A05, 35B50, 52A40

1. Introduction. The purpose of this paper is to show how convexity properties lead to sharp estimates for geometric quantities related to level curves of solutions to some classical partial differential equations.

We denote by $L(t)$ and $a(t)$ the perimeter and the area, respectively, of the domain bounded by a closed level curve $\{u=t\}\equiv\{x\colon u(x)=t\}$, where u is a real function on a plane domain.

We start by considering the solution u of the following capacity problem in a convex plane ring $D=D_0-\bar{D}_1$, where $D_1\subset D_0$ and D_0 and D_1 are plane convex domains:

$$\begin{aligned}&\Delta u=0 \quad \text{in } D,\\ &u=t_0 \quad \text{on } \partial D_0, \qquad u=t_1 \quad \text{on } \partial D_1,\end{aligned} \tag{1.1}$$

with

$$t_0 \text{ and } t_1 \text{ real constants.} \tag{1.2}$$

We derive isoperimetric inequalities involving the geometric quantities $L(t)$ and $a(t)$ related to the level curve $\{u=t\}$.

In Theorem 3.1 we prove that

(i) $\log L(t)$ *is a convex function of* t, and

(ii) $\log|a'(t)|$ *is a convex function of* t.

In the sequel ϕ is called logarithmic convex (concave) if $\log|\phi|$ is convex (concave).

In Theorems 4.1 and 4.2 we derive from (i) and (ii) further inequalities for the level sets of Green's function g for the Laplacian in a plane convex domain D. More specifically, if $\mu(t)$ is the distribution function of g, i.e., the area of the level set $\{g\geqq t\}$, we prove in Theorem 4.1 that:

(iii) $\log\mu(t)$ *is a convex function of* t.

In Theorem 4.2 a sharp upper bound for the length $L(t)$ of the level curves of g is obtained, namely, if L_0 is the perimeter of D:

(iv) $L(t)\leqq L_0\exp(-2\pi t)$.

In Theorem 3.2 we establish properties similar to (i)–(ii) for the following problem:

$$\begin{aligned}&\operatorname{div}(|\nabla u|^{p-2}\nabla u)=0 \quad \text{in } D,\\ &u=t_0 \quad \text{on } \partial D_0, \quad u=t_1 \quad \text{on } \partial D_1, \quad p>1.\end{aligned} \tag{1.3}$$

* Received by the editors October 21, 1985; accepted for publication (in revised form) March 24, 1987. This research was carried out while the author was visiting Cornell University, Ithaca, New York 14853.

† Istituto Analisi Globale e Applicazioni, via S. Marta 13/A, 50139 Firenze, Italy.

In Theorem 3.3 we give an explicitly sharp differential inequality for the length $L(t)$ of the level curves of the solution u to the following problem:

$$\begin{aligned} &\Delta u = f(u) \quad \text{in } D, \\ &u = t_0 \quad \text{on } \partial D_0, \quad u = t_1 \quad \text{on } \partial D_1, \quad t_0 < t_1, \\ &\text{with } f \geqq 0, \text{ and } f \text{ monotone nondecreasing.} \end{aligned} \tag{1.4}$$

In Theorem 3.4 we derive a sharp upper bound for the function $L(t)$ related to the solutions of problems (1.3) and (1.4). Furthermore in Theorem 3.5 we derive a sharp upper bound for the area $a(t)$ related to the solutions of problems (1.1), (1.3) in the case where D_1 is a circle.

Finally in § 5 we derive isoperimetric inequalities for the capacity problems (1.1)-(1.4) in the case where u satisfies the Bernoulli condition on the outside boundary ∂D_0, i.e.,

$$|\nabla u| = \text{const} > 0 \quad \text{on } \partial D_0. \tag{1.5}$$

Usually classical isoperimetric inequalities for the distribution function of Green's function or of solutions to elliptic equations are established by analysis and symmetrization arguments on the level sets (see [2]). Here we use arguments which stress the convexity properties of the function u. In fact in [5] and [7] it is proved, for an arbitrary dimension, that the solution u to (1.3) or (1.4), respectively, has convex level surfaces. An improvement for the problem (1.4) in dimension two is given in [4]. Strict convexity properties of Green's function g in a plane convex domain are shown in [4].

An interesting approach to a plasma physics problem [6] shows convexity properties of the level lines of harmonic functions which imply differential inequalities which supplement (i) and (ii).

The principal idea in the present paper is the introduction of a special coordinate system related to the convex level curves of u. More precisely, we consider the coordinates (θ, t) where t is the "level" of the curves $\{u = t\}$ and $(\cos\theta, \sin\theta)$ is the direction of the exterior normal vector to the level curve $\{u = t\}$. Furthermore we consider also the support function h of any level curve $\{u = t\}$ and rewrite any geometric quantity such as $L(t)$, $a(t)$, $|\nabla u|$ and the curvature K of $\{u = t\}$, in terms of h and its derivatives with respect to the curve parameters (θ, t). When u satisfies an elliptic differential equation, calculus arguments show that h is a solution to a nonlinear elliptic equation in (θ, t) coordinates. By analyzing this equation we obtain the sharp differential inequalities for $L(t)$ and $a(t)$, corresponding to (i)-(iv).

2. Support function. In this paragraph we start by recalling the geometric definition and the principal properties of the support function h of a plane convex domain D. For the necessary proofs and other details we refer the reader to [3]. Next we consider a family of support functions associated with a family of convex level curves of a given function u and show how certain geometric properties are defined in terms of derivatives of h.

Let D be a plane convex domain, and let us choose the origin of the coordinates inside D. Let us consider the exterior normal vector to ∂D at (x_1, x_2) given by $n = (\cos\theta, \sin\theta)$, for $\theta \in S \equiv [0, 2\pi)$. The distance from the origin to the support line r supporting ∂D at (x_1, x_2) orthogonal to n is given by the *support function*

$$h(\theta) = x_1 \cos\theta + x_2 \sin\theta. \tag{2.1}$$

If D is strictly convex r supports ∂D at only one point (x_1, x_2), h is of class $\mathscr{C}^1$ and the derivative of h with respect to θ is given by

$$h'(\theta) = -x_1 \sin\theta + x_2 \cos\theta. \tag{2.2}$$

If ∂D is $\mathscr{C}^2$ and has strictly positive curvature, then h is of class $\mathscr{C}^2$ also and

$$h(\theta) + h''(\theta) = R(\theta) > 0, \tag{2.3}$$

where $R(\theta)$ is the radius of curvature of ∂D.

For the proofs of the previous statements we refer the reader to [3, p. 18] and to [9, p. 3] where the formulas

$$L = \int_s R(\theta)\, d\theta = \int_s h(\theta)\, d\theta, \tag{2.4}$$

$$A = \frac{1}{2}\int_s h(\theta) R(\theta)\, d\theta, \tag{2.5}$$

for the perimeter L and the area A of D, respectively, also appear.

Let us now consider a real function u with strictly convex level curves in a domain D. Let us suppose also that u is of class $\mathscr{C}^2$ and that the derivative u_n of u along the outward normal to the level curve $\{u = t\}$ does not vanish at any point in D. For any value t in the range of u we consider the corresponding level curve of u: $\gamma_t \equiv \{u = t\}$, and for fixed t let $h(\theta, t)$ be the support function of the convex domain D_t bounded by γ_t. Furthermore, let $L(t)$, $a(t)$ and $R(\cdot, t)$ be the perimeter, the area and the radius of curvature of D_t, respectively. Partial derivatives are denoted by subscripts.

Of course by (2.1)-(2.3) the following equalities hold:

$$h(\theta, t) = x_1 \cos\theta + x_2 \sin\theta, \qquad (x_1, x_2) \in \gamma_t, \tag{2.6}$$

$$h_\theta(\theta, t) = -x_1 \sin\theta + x_2 \cos\theta, \qquad (x_1, x_2) \in \gamma_t, \tag{2.7}$$

$$h(\theta, t) + h_{\theta\theta}(\theta, t) = R(\theta, t) > 0 \tag{2.8}$$

where (x_1, x_2) is the unique point on γ_t with normal exterior vector $(\cos\theta, \sin\theta)$. Conversely, for any point (x_1, x_2) in D we can find θ and t by using,

$$(\cos\theta, \sin\theta) = \pm \nabla u(x_1, x_2)/|\nabla u(x_1, x_2)|, \tag{2.9}$$

$$t = u(x_1, x_2). \tag{2.10}$$

The + or − sign in (2.9) are given by the sign u_n. Moreover by (2.4) and (2.5) we have

$$L(t) = \int_s R(\theta, t)\, d\theta = \int_s h(\theta, t)\, d\theta, \tag{2.11}$$

$$a(t) = \frac{1}{2}\int_s h(\theta, t) R(\theta, t)\, d\theta. \tag{2.12}$$

We now show that certain classical expressions involving the partial derivatives of u with respect to x_1 and x_2 can be rewritten as derivatives of h with respect to θ and t. More precisely we have the following.

PROPOSITION 2.1. *If u has strictly convex level curves and its normal derivative u_n does not vanish at any point on D, then*

$$u_n = (h_t)^{-1}, \tag{2.13}$$

$$\Delta u = [-h_{tt} + (h_{\theta t}^2 + h_t^2) R^{-1}] h_t^{-3}. \tag{2.14}$$

Proof. We differentiate (2.6) and (2.10) with respect to t to get

$$h_t = \frac{\partial x_1}{\partial t}\cos\theta + \frac{\partial x_2}{\partial t}\sin\theta,$$

$$1 = u_{x_1}\frac{\partial x_1}{\partial t} + u_{x_2}\frac{\partial x_2}{\partial t};$$

expression (2.13) then follows with the use of (2.9). On the other hand, by differentiating (2.7) with respect to the exterior normal direction $n = (\cos\theta, \sin\theta)$, we get

$$h_{\theta t}\cdot t_n + h_{\theta\theta}\cdot\theta_n = -\frac{\partial x_1}{\partial n}\sin\theta - x_1\cos\theta\cdot\theta_n + \frac{\partial x_2}{\partial n}\cos\theta - x_2\sin\theta\cdot\theta_n. \tag{2.15}$$

By using (2.6), (2.13) and the fact that

$$t_n = u_n, \quad \frac{\partial x_1}{\partial n} = \cos\theta, \quad \frac{\partial x_2}{\partial n} = \sin\theta, \tag{2.16}$$

we can rewrite (2.15) in the form

$$h_{\theta t}\cdot h_t^{-1} + h_{\theta\theta}\cdot\theta_n = -h\cdot\theta_n.$$

So by solving for θ_n and using (2.8) we derive that the curvature of the orthogonal trajectories is given by

$$\theta_n = -h_{\theta t}h_t^{-1}R^{-1}. \tag{2.17}$$

Differentiating (2.13) with respect to n yields

$$U_{nn} = -(h_{tt}t_n + h_{\theta t}\theta_n)h_t^{-2}, \tag{2.18}$$

and so from (2.13) and (2.17) it follows that

$$U_{nn} = (-h_{tt} + h_{\theta t}^2 R^{-1})h_t^{-3}. \tag{2.19}$$

Now (2.14) follows from (2.13) and the following expression for the Laplacian in terms of normal derivatives of u and of the curvature $K = R^{-1}$ of the level curves of u:

$$\Delta u = u_{nn} + Ku_n. \qquad \square$$

3. Capacity functions in convex rings. We start by considering a harmonic function u with closed convex level curves $\gamma_t = \{x: u(x) = t\}$ in a convex ring D.

THEOREM 3.1. *If u is a solution to* (1.1)-(1.2), *then $L(t)$ is a logarithmic convex function in t, i.e.,*

$$L''L - (L')^2 \geqq 0, \tag{3.1}$$

and $|a'(t)|$ is a logarithmic convex function in t, i.e.,

$$a'''a' - (a'')^2 \geqq 0. \tag{3.2}$$

Moreover, equality holds in (3.1) *or* (3.2) *for some t if and only if all the level curves of u in D are concentric circles.*

Proof. Under the assumption that D_0 and D_1 are two convex domains bounding the convex level curves, $\{u = t_0\}$ and $\{u = t_1\}$, J. Lewis (cf. [7]) has proved that the function u has the following properties:

(a) $\{u = t\}$ is a strictly convex curve in D,

(b) $|\nabla u| \neq 0$ in D.

Therefore we can consider as in the previous section the support function $h(\theta, t)$ of the level curves of u.

We can suppose that $t_0 < t_1$, for substituting $-u$ for u leaves the inequalities (3.1) and (3.2) unchanged.

By (2.14) the following equality holds:

$$h_{tt} = (h_t^2 + h_{t\theta}^2) R^{-1} - h_t^{+3} \Delta u. \tag{3.3}$$

Since u is harmonic, we derive that

$$h_{tt} \geqq h_t^2 R^{-1}. \tag{3.4}$$

Moreover, by differentiating (2.11) with respect to t we get

$$L'(t) = \int_s h_t(\theta, t)\, d\theta, \tag{3.5}$$

$$L''(t) = \int_s h_{tt}(\theta, t)\, d\theta. \tag{3.6}$$

So from (3.4) it follows that

$$L''(t) \geqq \int_s h_t^2(\theta, t) R^{-1}(\theta, t)\, d\theta. \tag{3.7}$$

By Schwarz's inequality we have

$$\left(\int_s h_t\, d\theta \right)^2 \leqq \left(\int_s h_t^2 R^{-1}\, d\theta \right) \cdot \left(\int_s R\, d\theta \right). \tag{3.8}$$

So from (3.5), (3.6) and (2.11) we obtain (3.1). Equality in (3.1) holds for some $\tau \in (t_0, t_1)$ if and only if equality holds in (3.4) and (3.8) for $t = \tau$. This implies that

$$\begin{aligned} &h_{\theta t}(\cdot, \tau) \equiv 0 \quad \textit{on } S, \\ &h_t(\cdot, \tau) \textit{ is proportional to } R(\cdot, \tau) \textit{ on } S. \end{aligned} \tag{3.9}$$

Equivalently, we can say that

$$h_t(\cdot, \tau) \textit{ and } R(\cdot, \tau) \textit{ are constant on } S. \tag{3.10}$$

From (3.10) it then follows that $\{u = \tau\}$ is a circle and $|\nabla u| = |h_t|^{-1}$ is constant on $\{u = \tau\}$. From unique analytic continuation arguments it follows that any level curve of u is a circle concentric to D_τ.

Now we establish (3.2). By differentiating (2.8) and (2.12) with respect to t, we have

$$R_t = h_t + h_{\theta\theta t},$$

and

$$a'(t) = \frac{1}{2} \int_s (h_t R + h R_t)\, d\theta,$$

respectively. Replacing R_t we get

$$a'(t) = \frac{1}{2} \int_s (h_t R + h h_t + h h_{\theta\theta t})\, d\theta.$$

Integrating the last term in the previous integral two times by parts with respect to θ and using (2.8) yields

$$a'(t) = \int_s h_t R\, d\theta. \tag{3.11}$$

By differentiating (3.11) it follows that

$$a''(t) = \int_s (h_t R_t + R h_{tt})\, d\theta, \tag{3.12}$$

and using (3.3) it turns out that

$$a''(t) = \int_s (h_t h_{t\theta\theta} + 2h_t^2 + h_{t\theta}^2)\, d\theta - \int_s (R h_t^3\, \Delta u)\, d\theta.$$

Integrating the first term in the integral above by parts, we have

$$a''(t) = 2\int_s h_t^2\, d\theta - \int_s (R h_t^3\, \Delta u)\, d\theta \tag{3.13}$$

and so having used the fact that $\Delta u = 0$ we derive

$$a'''(t) = 4\int_s h_t h_{tt}\, d\theta. \tag{3.14}$$

Let us now observe that from (2.13) we have $h_t < 0$; therefore, from (3.4) and (3.14), it follows that

$$a'''(t) \leqq 4\int_s h_t^3 R^{-1}\, d\theta. \tag{3.15}$$

But Schwarz's inequality implies

$$\left(\int_s h_t^2\, d\theta\right)^2 \leqq \left(\int_s |h_t^3| R^{-1}\, d\theta\right) \cdot \left(\int_s |h_t| R\, d\theta\right). \tag{3.16}$$

So putting together (3.11), (3.13), (3.15), (3.16) we obtain (3.2). Moreover equality holds in (3.2) if and only if equality holds in (3.16) and (3.4), i.e., if (3.10) holds for some τ. The same previous analytic continuation argument completes the proof. □

The arguments of the previous theorem can also be applied to the solution u of the capacity problems (1.3) to obtain the following theorem.

THEOREM 3.2. *If u is a solution of* (1.3), *then the length $L(t)$ satisfies*

$$L \cdot L'' - \frac{1}{p-1}(L')^2 \geqq 0, \tag{3.17}$$

i.e., $(1/\alpha)L^\alpha$ is a convex function for $\alpha = (p-2)/(p-1)$, $p \neq 2$. Moreover the function $a(t)$ satisfies

$$a''' a' - \frac{2}{p}(a'')^2 \geqq 0, \tag{3.18}$$

i.e., $(1/\beta)|a'|^\beta$ is a convex function for $\beta = (p-2)/p$, $p \neq 2$. Equality holds in (3.17) *or* (3.18) *for some t if and only if all the level curves of u are concentric circles.*

Proof. Using normal coordinates, (1.3) becomes

$$\Delta u + (p-2)u_{nn} = 0. \tag{3.19}$$

So by (2.14) and (2.19) we have that

$$h_{tt} = \left(\frac{1}{p-1}h_t^2 + h_{\theta t}^2\right) R^{-1}, \tag{3.20}$$

from which it follows that

$$h_{tt} \geqq \frac{1}{p-1} h_t^2 R^{-1}. \tag{3.21}$$

If we now replace (3.4) by (3.21), then arguments along the lines of the previous theorem establish (3.17) and (3.18). □

Remark. By (3.3) the inequality (3.4) holds for any subharmonic function u with convex level curves and negative exterior normal derivative. So under this assumption the inequality (3.1) holds; moreover, by (3.12) and (3.3) one can show that $a'' > 0$.

Let us now consider the solution u to (1.4).

THEOREM 3.3. *If u is a solution to* (1.4), *then*

$$[\log L(t)]'' \geqq \frac{1}{4\pi^2} \cdot f(t) \cdot \frac{|L'(t)|^3}{L(t)}. \tag{3.22}$$

Equality in (3.22) *holds for some τ if and only if $\{u = \tau\}$ is a circle and $|\nabla u|$ is constant on $\{u = \tau\}$.*

Proof. Caffarelli and Friedman (cf. [4]) proved that u has strictly convex level curves. Moreover, $u_n = h_t^{-1} < 0$ since $t_0 < t_1$. So by (3.3) we get

$$h_{tt} \geqq h_t^2 R^{-1} + f(t)|h_t|^{+3}.$$

By integrating the above inequality on S and by using Holder's inequality for the last term, the same arguments in the proof of (3.1) prove (3.22). □

We now show that (3.17) (respectively (3.22)) leads to a sharp upper bound for the length $L(t)$ of the level curves of the solution u to (1.3) (respectively (1.4)).

Let $\tilde{D}_0$ and $\tilde{D}_1$ be two concentric circles with the same perimeters as D_0 and D_1. We consider the radial solutions v to (1.3) or to (1.4) with the following boundary conditions:

$$v = t_0 \quad \text{on } \partial\tilde{D}_0, \qquad v = t_1 \text{ on } \partial\tilde{D}_1. \tag{3.23}$$

We call v the *L-symmetrization* of u in $D_0 - D_1$.

THEOREM 3.4. *If u is a solution of* (1.3) *or* (1.4) *satisfying* (1.2) *and v is the L-symmetrization of u, then*

$$L(t) \leqq l(t) \tag{3.24}$$

where $l(t)$ is the perimeter of the level circles $\{v = t\}$.

Proof. We first consider the problem (1.3). By (3.23) we have

$$l(t_0) = L(t_0), \qquad l(t_1) = L(t_1).$$

Moreover by Theorem 3.2 we get that $l(t)$ satisfies

$$l(t) \cdot l''(t) - \frac{1}{p-1} l'(t)^2 = 0, \qquad t \in (t_0, t_1). \tag{3.25}$$

Similarly for the problem (1.4) we get

$$[\log l(t)]'' = \frac{1}{4\pi} f(t) \cdot \frac{|l'(t)|^3}{l(t)}, \qquad t \in (t_0, t_1). \tag{3.26}$$

By comparing (3.25) and (3.26) with (3.17) and (3.22), respectively, and by standard comparison theorems for differential inequalities (cf. [8]), we prove (3.24). □

We now wish to establish an upper bound for the area $a(t)$ in the problem (1.31). So let D_0^* and D_1^* be two concentric circles with the same area as D_0 and D_1. Let w be the radial solution to (1.3), with the following boundary conditions:

$$w = t_0 \quad \text{on } \partial D_0^*, \qquad w = t_1 \quad \text{on } \partial D_1^*.$$

We call w the *a-symmetrization of u in* $D_0 - D_1$.

THEOREM 3.5. *Let u be a solution of* (1.3) *and let w be the a-symmetrization of u. If*

(3.27) $$D_1 \text{ is a circle,}$$

then for any $p > 1$

(3.28) $$a''a - \frac{p}{2(p-1)}(a')^2 \geqq 0$$

and

(3.29) $$a(t) \leqq A(t),$$

where $A(t)$ *is the area of the level circle* $\{w = t\}$.

Proof. First we prove (3.28) and (3.29) for $p = 2$. Let us set

(3.30) $$M(t) = \frac{a''(t) \cdot a(t) - (a'(t))^2}{a(t)}.$$

By differentiating, (3.30) becomes

$$M' \equiv a''' - 2\frac{a'a''}{a} + \frac{(a')^3}{a^2}.$$

Since $a' < 0$, we get from (3.2) that

$$a''' \leqq \frac{(a'')^2}{a'}$$

and so

$$M' \leqq \frac{(a'')^2}{a'} - 2\frac{a'a''}{a} + \frac{(a')^3}{a^2}$$

$$= \frac{a''}{a}\left(\frac{a''a - (a')^2}{a'}\right) + \frac{a'}{a}\left(\frac{(a')^2 - a''a}{a}\right) \equiv \frac{M^2}{a'},$$

from which it follows that

(3.31) $$M'(t) \leqq 0.$$

Let us now compute $M(t_1)$. By (2.12), (3.11) and (3.13) we get

$$a \cdot M \equiv \left(\int_s h_t^2 \, d\theta\right) \cdot \left(\int_s hR \, d\theta\right) - \left(\int_s h_t R \, d\theta\right)^2.$$

By assumption (3.27) we have $R(\cdot, t_1) \equiv$ constant, say R_1, and so

$$\frac{aM(t_1)}{R_1^2} = 2\pi\left(\int_s h_t^2(\theta, t_1) \, d\theta\right) - \left(\int_s h_t(\theta, t_1) \, d\theta\right)^2.$$

Schwarz's inequality implies that

$$M(t_1) \geqq 0.$$

So by (3.31) we get $M(t) \geqq 0$, and (3.28) follows from (3.30) for $p=2$; by standard comparison theorems (cf. [8]), (3.28) implies (3.29).

For $p>1$, $p \neq 2$, by replacing (3.30) with

$$M \equiv \left[a''a - \frac{p}{2(p-1)}(a')^2 \right] a^{-\alpha}$$

where $\alpha = 1/(p-1)$, we can establish (3.28) and (3.29) by using arguments similar to those for the case $p=2$. □

Remark. In Theorem 5.2 we prove that for $p=2$, in the case where (3.27) is replaced by the Bernoulli condition (1.5), the opposite inequality of (3.28) holds. So in general assumption (3.27) is essential for inequalities (3.28) and (3.29) to hold.

4. Isoperimetric inequalities for Green's function. Usually the point of departure in level set-analysis of a function u is the coarea formula (cf. [2, p. 52]):

$$|\mu'(t)| = \oint_{u=t} \frac{ds}{|\nabla u|}, \tag{4.1}$$

where $\mu(t)$ is the distribution function of u, i.e., the Lebesgue measure of the level set $\{u \geqq t\}$. Moreover, applying Schwarz's inequality to (3.25), the following inequality is usually considered:

$$|\mu'(t)| \geqq \left(\oint_{u=t} ds \right)^2 \Big/ \oint_{u=t} |\nabla u| \, ds. \tag{4.2}$$

Let us now consider Green's functions of the Laplace operator in a plane domain D. It is of the form

$$g(x, y) = \frac{1}{2\pi} \log \frac{R(y)}{|x-y|} + H(x, y), \tag{4.3}$$

where H is determined such that for fixed $y \in D$ we have the following:

(i) $g(x, y) = 0$ for $x \in \partial D$,

(ii) $H(\cdot, y)$ is harmonic in D and continuous in $\bar{D}$,

(iii) $H(y, y) = 0$.

$R(y)$ is called the *conformal radius* of D with respect to y. Let us define D^* as the circle with the same area as D and with center at the origin of the coordinates and radius R^*. Let D_y be a concentric circle to D^*, with radius $R_y = R(y)$. We consider the functions g^* and g_y given by

$$\begin{aligned} g^*(x) &= \frac{1}{2\pi} \log \frac{R^*}{|x|}, \\ g_y(x) &= \frac{1}{2\pi} \log \frac{R_y}{|x|}, \end{aligned} \tag{4.4}$$

respectively. For fixed y, we denote the distribution functions of $g(\cdot, y)$, g^*, g_y, by $\mu(t)$, $\mu_*(t)$ and $\mu_y(t)$, respectively. By (4.3), (4.4) it follows that

$$\begin{aligned} \mu(t) &= \pi (R_y)^2 \exp(-4\pi t) + o(\exp(-4\pi t)), \\ \mu_*(t) &= \pi (R^*)^2 \exp(-4\pi t), \\ \mu_y(t) &= \pi (R_y)^2 \exp(-4\pi t). \end{aligned} \tag{4.5}$$

Since $\oint |\nabla g| \, ds = 1$ on any level curve of g, it follows from (4.2) that

$$|\mu'(t)| \geqq L^2(t). \tag{4.6}$$

From the classical isoperimetric inequality we derive

$$|\mu'(t)| \geqq 4\pi\mu(t), \tag{4.7}$$

and by integration (see also [2])

$$\mu_y(t) \leqq \mu(t) \leqq \mu_*(t). \tag{4.8}$$

If D is a convex plane domain the following theorem is an improvement of the inequality (4.7).

THEOREM 4.1. *Let D be a convex plane domain and let y be a fixed point in D, then the following differential inequalities hold*:

$$\mu'''\mu' - (\mu'')^2 \geqq 0, \tag{4.9}$$

i.e., $\log|\mu'|$ *is a convex function*;

$$\mu''\mu - (\mu')^2 \geqq 0, \tag{4.10}$$

i.e., $\log\mu$ *is a convex function*; *and*

$$\mu'' \geqq 4\pi|\mu'|. \tag{4.11}$$

Moreover D is a circle with center at y if and only if equality holds in (4.9), (4.10) *or* (4.11) *for some t.*

Proof. In [4] the authors prove that g has strictly convex level curves, so (4.9) follows directly from Theorem 3.1. To prove (4.10) we can repeat similar arguments to those in the proof of (3.28). Another proof of (4.10) follows by considering a sequence $u^{(m)}$ of harmonic functions in $D - D^{(m)}$, satisfying

$$u^{(m)} = 0 \quad \text{on } \partial D, \qquad u^{(m)} = C^{(m)} \quad \text{on } \partial D^{(m)},$$

where $D^{(m)}$ is a circle with center at y and such that

$$\oint_{\partial D^{(m)}} |\nabla u^{(m)}|\, ds = 1.$$

Indeed it is easy to see that $u^{(m)}$ approaches g when $C^{(m)} \to \infty$. Now by applying Theorem 3.5 to $u^{(m)}$ we get that any distribution function $a^{(m)}$ of $u^{(m)}$ is logarithmic convex. Since $\mu = \lim_{m\to\infty} a^{(m)}$, we derive that μ is also logarithmic convex. This proves (4.10).

Finally (4.11) follows from (4.7) and (4.10). This concludes the proof. □

The following corollary follows directly from (4.10), by applying standard comparison theorems for differential inequalities (cf. [8]).

COROLLARY 4.1. *Let $0 \leqq t_0 < t_1$ be fixed constants, and y a fixed point in D. Let D_0 and D_1 be the level sets $\{g \geqq t_0\}$ and $\{g \geqq t_1\}$, respectively. If w is the a-symmetrization of g in $D_0 - D_1$, then*

$$\mu(t) \leqq A(t) \quad \text{for } t \in (t_0, t_1), \tag{4.12}$$

where $A(t)$ is the area of the circle $\{w = t\}$.

We now wish to establish a lower and an upper bound for the length $L(t)$ of the level curves $\{g = t\}$.

Of course from the left-hand inequality in (4.8) and the classic isoperimetric inequality we get

$$2\pi R_y \exp(-2\pi t) \leqq L(t). \tag{4.13}$$

Unfortunately a similar argument does not apply to the right-hand inequality in (4.9) which will give an upper bound for $L(t)$. However, in the following theorem we give an upper bound for $L(t)$ which only depends on the perimeter L_0 of D.

THEOREM 4.2. *Let D be a convex plane domain with perimeter L_0 and let y be a fixed point in D. Then L is logarithmic convex and*

$$L(t) \leqq L_0 \exp(-2\pi t). \tag{4.14}$$

Equality holds in (4.14) *if and only if D is a circle centered at y.*

Proof. As in the proof of the previous theorem we have that g has strictly convex level curves. So Theorem 3.1 applies and $\log L$ is convex. Moreover, by (4.8) it follows that

$$\log \mu_y \leqq \log \mu \leqq \log \mu_*.$$

Since $\log \mu$ is convex and $\log \mu_y$, $\log \mu_*$ are linear functions with slope -4π, it follows that

$$\lim_{t\to+\infty} \frac{\mu'(t)}{\mu(t)} = -4\pi.$$

From (4.5) and the previous equality we have that

$$\lim_{t\to+\infty} \frac{|\mu'(t)|}{\exp(-4\pi t)} = 4\pi^2 R_y^2,$$

and so from (4.6),

$$\lim_{t\to+\infty} \frac{L(t)}{\exp(-2\pi t)} \leqq 2\pi R_y.$$

By logarithmic convexity properties of L and the previous equality we derive (4.14). □

5. Optimal conductors. In [10] the following result is proved: given a convex domain D_1 and a constant $\gamma > 0$ there exists a unique convex domain $D_0 \supset D_1$, such that the solution u to (1.1), (1.2) satisfies the Bernoulli condition:

$$|\nabla u| = \gamma \quad \text{on } \partial D_0. \tag{5.1}$$

This problem arises in optimal conductors and in some classic free boundary problems (see [1]).

The results of § 3 can be applied to obtain isoperimetric inequalities for the *optimal conductor* $D_0 - D_1$. For simplicity, let $t_0 = 0$, $t_1 = 1$ in (1.2). So the logarithmic capacity of $D_0 - D_1$ is given by

$$C = \oint_{\partial D_0} |\nabla u| \, ds, \tag{5.2}$$

and the constant γ in (5.1) is given by

$$\gamma = C/L_0, \tag{5.3}$$

where L_0 is the perimeter of ∂D_0.

THEOREM 5.1. *Let L_0, L_1 be the perimeter, and A_0, A_1 be the area of D_0 and D_1, respectively.*

The logarithmic capacity C of the optimal conductor $D_0 - D_1$ satisfies the following inequalities:

$$\frac{1}{C} \geqq \frac{1}{2\pi} \log \frac{L_0}{L_1}, \tag{5.4}$$

$$\frac{1}{C} \leqq \frac{1}{4\pi} \log \left[\frac{L_0^2}{L_0^2 - 4\pi(A_0 - A_1)} \right]. \tag{5.5}$$

Equality holds in (5.4) *or* (5.5) *if and only if $D_0 - D_1$ is a circular annulus.*

Proof. By logarithmic convexity of the function $L(t)$, Theorem 3.1, we have

$$\log L(1) \geqq \log L(0) + \frac{L'(0)}{L(0)}. \tag{5.6}$$

Moreover by (3.5), (5.1)-(5.3) we have

$$\frac{L'(0)}{L(0)} = -\frac{2\pi}{C}. \tag{5.7}$$

Now (5.4) follows by (5.6) and (5.7).

Similarly by Theorem 3.2 we derive

$$\log |a'(t)| \geqq \log |a'(0)| + \frac{a''(0)}{a'(0)} \cdot t \tag{5.8}$$

and from (3.11), (3.12) we have

$$|a'(0)| = \frac{L_0^2}{C}, \qquad a''(0) = \frac{4\pi L_0^2}{C^2}. \tag{5.9}$$

So by integrating (5.8) and by the equalities above, we get

$$A_0 - A_1 \geqq \frac{L_0^2}{4\pi}\left[1 - \exp\left(-\frac{4\pi}{C}\right)\right].$$

Inequality (5.5) is proved now by solving for $1/C$ in the inequality above. □

Remark 5.1. The length L_0 in (5.4) and (5.5) is not explicitly given, since L_0 is the length of the free boundary ∂D_0.

Moreover, by computation one can show that the upper bound given in (5.5) for $1/C$ is less than in Carleman's inequality:

$$\frac{1}{C} \leqq \frac{1}{4\pi} \log \frac{A_0}{A_1}. \tag{5.10}$$

In the following theorem we give an explicit isoperimetric inequality for the free boundary ∂D_0.

THEOREM 5.2. *If u is a solution to* (1.1), (1.2), *satisfying* (5.1), *then*

$$L_0^2 - 4\pi A_0 \leqq L_1^2 - 4\pi A_1. \tag{5.11}$$

Moreover, the following differential inequalities hold:

$$(L^2 - 4\pi a)' \geqq 0, \tag{5.12}$$

$$a''a - (a')^2 \leqq 0, \tag{5.13}$$

i.e., a is a logarithmic concave function. Equality holds in (5.11), (5.12) *or* (5.13) *if and only if D_1 is a circle.*

Proof. We obtain (5.11) by comparing the two terms on the right-hand side of (5.4) and (5.5). More simply, (5.11) follows by (5.12) which we now prove. In fact, let us set

$$G(t) \equiv L^2(t) - 4\pi a(t). \tag{5.14}$$

Since $\log L$ is convex (Theorem 3.1) we have that L'/L is increasing and by (5.7)

$$L'(t) \geqq \frac{-2\pi}{C} L(t). \tag{5.15}$$

So by (5.14)

$$G'(t) \geqq -4\pi\left[\frac{L^2(t)}{C} + a'(t)\right].$$

Inequality (5.12) follows now from (4.2) and the inequality above.

Finally to prove (5.13) we consider the function M defined by (3.30). By (5.9) we have

$$M(0) = \left(4\pi\frac{L_0^2}{C^2}A_0 - \frac{L_0^4}{C^2}\right)\frac{1}{A_0}.$$

So by classical isoperimetric inequality $M(0) \leqq 0$ and by (3.31) we derive $M(t) \leqq 0$, which implies (5.13). □

Finally, let us observe that: for the solution u to (1.4), and satisfying (5.1), the inequalities (5.4) and (5.12) hold. (5.4) and (5.13) also can be extended in a suitable form to the problems (1.3) satisfying the Bernoulli condition (5.1).

Acknowledgments. The author wishes to thank Cornell University for its hospitality and Professor L. E. Payne for his advice and the several fruitful discussions about the problems presented here.

REFERENCES

[1] A. ACKER, *A free boundary optimisation problem*, this Journal, 9 (1978), pp. 1179-1191.
[2] C. BANDLE, *Isoperimetric Inequalities and their Applications*, Pitman, London, 1980.
[3] T. BONNESEN AND W. FENCHEL, *Theorie der Konvexen Korper*, Springer, Berlin, 1934.
[4] L. A. CAFFARELLI AND A. FRIEDMAN, *Convexity of solutions of semilinear elliptic equations*, Duke Math. J., 52 (1985), pp. 431-456.
[5] L. A. CAFFARELLI AND J. SPRUCK, *Convexity properties of solutions of some classic variational problems*, Comm. Partial Differential Equations, 7 (1982), pp. 1337-1379.
[6] P. LAURENCE AND E. STREDULINSKY, *A new approach to queer differential equations*, Comm. Pure Appl. Math., 38 (1985), pp. 333-355.
[7] J. LEWIS, *Capacitary functions in convex rings*, Arch. Rational Mech. Anal., 66 (1977), pp. 201-224.
[8] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[9] L. A. SANTALÒ, *Integral geometry and geometric probability*, Encyclopedia of Mathematics and Its Applications, vol. 1, Addison-Wesley, London, 1976.
[10] D. E. TEPPER, *Free boundary optimisation problem*, this Journal, 5 (1974), pp. 841-846.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

011

STARLIKE FUNCTIONS AND LINEAR FUNCTIONS OF A DIRICHLET DISTRIBUTED VECTOR*

JYH-MING JIANG†

Abstract. It is shown that under specific conditions the limiting distribution of $\sum_{i=1}^{k} u_i z_i$ will be a uniform distribution on the unit disk where $z_1, \cdots, z_k$ are increasingly dense points on the unit circle and $\mathbf{u} = (u_1, \cdots, u_k)$ has a Dirichlet distribution. More general circularly symmetric distributions are also obtained as such limits. This is motivated by a new representation of starlike functions as expectations over the unit disk. A new kind of characteristic function and its convergence theorem are used.

Key words. Dirichlet distribution, spherical distributions, Carlson's R, convergence theorem about d-transformation, d-characteristic function

AMS(MOS) subject classifications. primary 60E10; secondary 30C45

1. Introduction. The study of univalent and starlike functions of a complex variable provides a motivation in § 2 for the problem of the limiting distribution of a linear function of a Dirichlet vector. This problem is also of intrinsic interest in probability theory as the distribution of the empirical mean of a Dirichlet random process on the circle as index set. In § 3, we define new kinds of characteristic functions. We show that they have properties similar to the properties of the traditional characteristic function. We use these properties in § 4 to determine the limiting distribution of a linear function of a Dirichlet vector, under some regularity conditions, as the coefficients grow increasingly dense on the unit circle.

2. Motivation. Let S be the full class of functions that are analytic on the open unit disk, one-to-one and normalized (i.e., $f(\zeta) = \zeta + a_2\zeta^2 + a_3\zeta^3, \cdots$, where ζ is a complex variable, $f(0) = 0$, and $f'(0) = 1$). The recently proved theorem of Bieberbach (de Branges (1985)) states that if $f \in S$, then $|a_n| \leqq n$ for $n = 0, 1, 2, \cdots$. Further define f as a *starlike* function if and only if tw belongs to the range of f for any $t \in [0, 1]$ and any w in the range of f. If we let S_t be the set of starlike functions, then S_t is a subset of S. We have the following well-known representation theorem (see Schober (1975, Thm. 2.13, p. 12)).

THEOREM 2.1.

$$s \in S_t \Leftrightarrow s(\zeta) = \zeta \cdot \exp \int_0^{2\pi} -2 \log [1 - (\exp (i\theta))\zeta] \, d\nu(\theta) \tag{2.1}$$

where ν is some probability measure on $[0, 2\pi)$.

Consider the case that the measure ν is discrete (the corresponding functions s are dense in S_t), and assume that

$$\nu(\{\theta_j\}) = t_j, \quad 1 \leqq j \leqq k, \quad t_j > 0 \quad \text{and} \quad \sum_{j=1}^{k} t_j = 1.$$

We then have

$$\frac{s(\zeta)}{\zeta} = \prod_{j=1}^{k} (1 - z_j\zeta)^{-2t_j} \tag{2.2}$$

where $z_j = \exp (i\theta_j)$. We shall derive a new representation for such $s(\zeta)$.

* Received by the editors February 26, 1986; accepted for publication (in revised form) February 28, 1987. This paper is based on parts of the author's unpublished Ph.D. dissertation from the State University of New York, Albany, New York 12222.

† Department of Mathematics, University of Lowell, Lowell, Maine 01854. This research was supported in part by the National Science Foundation grants MCS-8301335 and DMS-8614793. This work was completed while the author was an assistant professor in the Mathematical Systems Program, Sangamon State University, Springfield, Illinois 62794-9243.

DEFINITIONS. The random vector $\mathbf{u}$ is said to have the Dirichlet distribution $D(\mathbf{b})$ with parameter $\mathbf{b}=(b_1, \cdots, b_k)'$, where every $b_i>0$, if $\mathbf{u}$ has the density in any $k-1$ of its coordinates,

$$f(\mathbf{u}; \mathbf{b}) = B^{-1}(\mathbf{b}) \cdot \prod_{i=1}^{k} u_i^{b_i-1}, \tag{2.3a}$$

for all $\mathbf{u}$ in the probability simplex $\{\mathbf{u}$: each $u_i \geqq 0$, $u_\cdot = 1\}$, where $u_\cdot = u_1 + \cdots + u_k$, and

$$B(\mathbf{b}) = \prod_{i=1}^{k} \Gamma(b_i)/\Gamma(b_\cdot).$$

Following Carlson (1977), define $R_n(\mathbf{b}, \mathbf{z})$ as the nth moment of the random variable $\theta = \sum_{j=1}^{k} u_j z_j$, where $\mathbf{u} \sim D(\mathbf{b})$,

$$R_n(\mathbf{b}, \mathbf{z}) = \underset{u|b}{E}\, \theta^n = \underset{u|b}{E} (\mathbf{u} \cdot \mathbf{z})^n. \tag{2.3b}$$

For a discussion of the relationship between Carlson's R functions and starlike functions see Carlson and Shaffer (1984).

LEMMA 2.2 (Carlson (1977, p. 143)). *Let $\mathbf{b} \in R^k$, $\mathbf{z} \in C^k$, $t \in C$ (R is the real line, C is the complex plane and R^k and C^k are the kth Cartesian power of R and C, respectively). Assume that $\max_{1 \leqq j \leqq k} |tz_j| < 1$. Then*

$$f_k(t) \equiv \prod_{j=1}^{k} (1 - itz_j)^{-b_j} = \sum_{n=0}^{\infty} i^n \cdot t^n \cdot \frac{(b_\cdot, n)}{n!} R_n(\mathbf{b}, \mathbf{z}), \tag{2.4}$$

using Appell's notation,

$$(b_\cdot, n) = \Gamma(b_\cdot + n)/\Gamma(b_\cdot) = b_\cdot(b_\cdot + 1) \cdots (b_\cdot + n - 1).$$

The following lemma can be proved by using the previous lemma and the following equation:

$$(1-y)^{-a} = \sum_{n=0}^{\infty} \frac{(a, n)}{n!} y^n, \tag{2.5}$$

for every real number a and $|y| < 1$.

LEMMA 2.3. *Let μ_b be a Dirichlet measure with parameter $\mathbf{b} = (2t_1, 2t_2, \cdots, 2t_k)'$, where t_j's > 0, and $\sum t_j = 1$. Then for any ζ in the unit disk, if $\mathbf{z} = (z_1, \cdots, z_k)'$ and the z_j's are points on the unit circle,*

$$\prod_{j=1}^{k} (1 - z_j\zeta)^{-2t_j} = \int_E [1 - (\mathbf{u} \cdot \mathbf{z})\zeta]^{-2}\, d\mu_b(\mathbf{u}), \tag{2.6}$$

where E is the probability simplex $\{\mathbf{u}$: each $u_i \geqq 0$, $u_\cdot = 1\}$.

By Lemma 2.3 and Theorem 2.1,

$$\zeta \cdot \int_E [1 - (\mathbf{u} \cdot \mathbf{z})\zeta]^{-2}\, d\mu_b(u) \tag{2.7}$$

is a starlike function for the case of discrete ν (2.1) and $b_\cdot = 2$. We remark that Bieberbach's theorem is straightforward for any function (2.7). This motivates our consideration of the following problems.

Our aim is to find the limiting distribution of $\mathbf{u}_{k*} \cdot \mathbf{z}_{k*} = \sum_{j=1}^{k} u_{kj} z_{kj}$ as $k \to \infty$, where $\mathbf{u}_{k*} \sim D(\mathbf{b}_{k*})$ and the z_{kj}'s are points on the unit circle. If the set of z_{kj}'s for specified k becomes dense on the circle as k increases, then our Dirichlet distribution has a

Dirichlet random process on the circle as limit in distribution (Ferguson (1973)). Thus the limiting distribution of our random sum is also of interest as the distribution of the empirical mean of a Dirichlet random process.

3. *d*-characteristic functions. Lord (1954) shows that a spherical distribution is determined by its marginal distribution. We shall show that under specific conditions, the limiting distribution of $\mathbf{u}_{k*} \cdot \mathbf{z}_{k*}$ is spherically symmetric. Therefore, we need only to find the limiting distribution of the real part of $\mathbf{u}_{k*} \cdot \mathbf{z}_{k*}$ (that is, $\mathbf{u}_{k*} \cdot \mathbf{x}_{k*}$ where each $z_{kj} = x_{kj} + iy_{kj}$).

One traditional method of finding the limiting distribution of a sequence of random variables is to find the limit of the corresponding (traditional) characteristic functions. But this method seems overly complicated for dealing with our problems. Therefore, we use the following alternative characteristic functions. For convenience, we shall call these *d*-characteristic functions.

Let

$$g(t;\, W, d) = \underset{W}{E}[(1 - itW)^{-d}], \qquad |t| < 1, \quad d > 0, \tag{3.1}$$

where W is any random variable on $[-1, 1]$. More generally, for any finite measure μ with supports in $[-a, a]$, we define the *d-transformation* of μ as

$$\hat{\mu}^d(t) = \int_{-a}^{a} (1 - itx)^{-d}\, d\mu(x), \qquad |t| < a^{-1}, \quad d > 0, \tag{3.2}$$

where a is a positive real number. We will show there is a one-to-one correspondence between $\hat{\mu}^d(t)$ and μ.

LEMMA 3.1. *For any finite measures μ and ν with supports in $[-a, a]$ and any positive real number d, if we have*

$$\hat{\mu}^d(t) = \hat{\nu}^d(t), \tag{3.3}$$

for all $|t| < a^{-1}$, then $\mu = \nu$.

Proof. Expand the integrands by (2.5), equation (3.3) is then equivalent to the following equation:

$$\sum_{n=0}^{\infty} \frac{(d, n)}{n!} i^n \cdot t^n \cdot \int_{-a}^{a} x^n\, d\mu(x) = \sum_{n=0}^{\infty} \frac{(d, n)}{n!} i^n \cdot t^n \cdot \int_{-a}^{a} x^n\, d\nu(x),$$

for all $|t| < a^{-1}$. If we regard t as variable, then after equating (for each n) the corresponding coefficients of t^n in the two sums, we have

$$\int_{-a}^{a} P(x)\, d\mu(x) = \int_{-a}^{a} P(x)\, d\nu(x), \tag{3.4}$$

where $P(x)$ is any polynomial function, and similarly for any continuous function. This implies that $\mu = \nu$.

LEMMA 3.2. *If* $\mathbf{u}_{k*} \sim D(\mathbf{b}_{k*})$, $\mathbf{u}_{k*} = (u_{k1}, \cdots, u_{kk})'$, $\mathbf{b}_{k*} = (b_{k1}, \cdots, b_{kk})'$, $b_{k\cdot} = \sum_{j=1}^{k} b_{kj}$ *and* $W_k = \mathbf{u}_{k*} \cdot \mathbf{x}_{k*}$, *then*

$$g(t; W_k, d) = R_{-d}(\mathbf{b}_{k*};\, 1 - itx_{k1}, \cdots, 1 - itx_{kk}) \tag{3.5}$$

where g is defined by (3.1).

Proof.

$$g(t; W_k, d) = \underset{u_{k*}|b_{k*}}{E}\left(1 - it\left(\sum_{j=1}^{k} u_{kj}x_{kj}\right)\right)^{-d}$$

$$= \underset{u_{k*}|b_{k*}}{E}\left(\sum_{j=1}^{k} u_{kj}(1 - itx_{kj})\right)^{-d}$$

$$= R_{-d}(\mathbf{b}_{k*}; 1 - itx_{k1}, \cdots, 1 - itx_{kk}).$$

The last identity follows from the definition (2.3b).

The following corollary can be shown by formula (6.6.5) in Carlson (1977).

COROLLARY 3.3. *Let* $c = b_{k\cdot} = \sum_{j=1}^{k} b_{kj}$, *then*

$$g(t; W_k, c) = \prod_{j=1}^{k} (1 - itx_{kj})^{-b_{kj}}. \tag{3.6}$$

We give the important convergence theorem about d-transformations. This is analogous to the corresponding convergence theorem for the Fourier transformations. Before we state the theorem, we need to give the following definitions and lemma.

DEFINITIONS. Let Ω be a measured space, μ is a measure on Ω. Then μ is called a *subprobability measure* (s.p.m.) if $\mu(\Omega) \leqq 1$. An interval (a, b) is called a *continuity interval* of μ if and only if $\mu(a, b) = \mu[a, b]$; in other words neither a nor b is an atom of μ. A sequence $\{\mu_n, n \geqq 1\}$ of s.p.m.'s is said to *converge vaguely* to an s.p.m. μ if and only if for every continuity interval $(a, b]$ of μ, we have $\mu_n(a, b] \to \mu(a, b]$. This will be denoted by $\mu_n \xrightarrow{v} \mu$.

LEMMA 3.4. *If* μ *and the* μ_n*'s are* s.p.m. *with supports in* $[-a, a]$ *and* $\mu_n \xrightarrow{v} \mu$, *then*

$$\hat{\mu}_n^d(t) \to \hat{\mu}^d(t) \quad \forall d > 0 \quad \textit{and} \quad |t| < a^{-1}, \tag{3.7}$$

where $\hat{\mu}$ *is defined by* (3.2).

Proof. By Theorem 4.4.1 of Chung (1974), we have

$$\int_R (1 - itx)^{-d}\, d\mu_n(x) \to \int_R (1 - itx)^{-d}\, d\mu(x) \quad \forall d > 0 \quad \text{and} \quad |t| < a^{-1}.$$

Given d, then by Lemma 3.1, subprobability measures have a one-to-one correspondence with their d-transforms.

Now, we are ready to give the following convergence theorem.

THEOREM 3.5. *Given* d, *assume the subprobability measures* $\mu, \mu_1, \mu_2, \cdots$ (*with supports in* $[-a, a]$) *correspond to* $\hat{\mu}^d(t), \hat{\mu}_1^d(t), \hat{\mu}_2^d(t), \cdots$, *respectively. If for all* $|t| < a^{-1}$

$$\hat{\mu}_n^d(t) \to \hat{\mu}^d(t) \quad \textit{as } n \to \infty, \tag{3.8}$$

then

$$\mu_n \xrightarrow{v} \mu \quad \textit{as } n \to \infty. \tag{3.9}$$

Proof. For any s.p.m. sequence $\{\mu_n\}$, by Theorem 4.3.3 of Chung (1974), there is a subsequence that converges vaguely to an s.p.m. say $\mu_{n_k} \xrightarrow{v} \lambda$, as $n_k \to \infty$. Then by Lemma 3.4, $\hat{\mu}_{n_k}^d(t) \to \hat{\lambda}^d(t)$, for all $|t| < a^{-1}$. By (3.8), we also have $\hat{\mu}_{n_k}^d(t) \to \hat{\mu}^d(t)$, for all $|t| < a^{-1}$. Therefore, we have $\hat{\lambda}^d(t) = \hat{\mu}^d(t)$, for all $|t| < a^{-1}$. By the uniqueness property of Lemma 3.1, we have $\lambda = \mu$. Therefore, we have $\mu_{n_k} \xrightarrow{v} \mu$, as $n_k \to \infty$. If there is another vaguely convergent subsequence of $\{\mu_n\}$ converging to ψ, then, by the same arguments as above, $\psi = \mu$. By Theorem 4.3.4 of Chung (1974), this proves $\mu_n \xrightarrow{v} \mu$.

4. Limiting distributions. Before proceeding with an application of Theorem 3.5, we need to have the following definitions and lemmas. Define a sequence of random variables $\{X_n\}$ to converge in distribution to X if and only if the sequence $\{\mu_n\}$ of corresponding probability measures converges vaguely to the probability measure μ of X.

Regularity conditions. Given $c>0$, $b=\{b_{kj}: k\geqq 2 \text{ and } 1\leqq j\leqq k\}$ and $z=\{z_{kj}: k\geqq 2 \text{ and } 1\leqq j\leqq k\}$. If b satisfies the following properties:

(A) $b_{k.}\equiv \sum_{j=1}^{k} b_{kj}=c \quad \forall k\geqq 2,$

(B) $B_k\equiv \max_{1\leqq j\leqq k} b_{kj}\to 0 \quad \text{as } k\to\infty,$

and z satisfies the following properties:

(C) z_{kj} lies on the unit circle with center at $(0,0)$,

(D) $\theta_{k,j-1}\leqq \arg z_{kj}<\theta_{kj}$, $1\leqq j\leqq k$ and for all $k\geqq 2$,

where $\theta_{k0}=0$, $\theta_{kj}=(2\pi/c)\sum_{m=1}^{j} b_{km}$ and $\arg z_{kj}$ denotes the argument of z_{kj}, then we say that b and z satisfy *the regularity conditions.* Notice that condition (B) is needed for the next lemma in the transition from a Riemann sum to an integral.

LEMMA 4.1. *If* $\mathbf{u}_{k*}\sim D(\mathbf{b}_{k*})$, $W_k^{\phi}=\mathbf{u}_{k*}\cdot \mathbf{z}_{k*}^{\phi}$ *and the regularity conditions* (A)-(D) *hold, where* $z_{kj}^{\phi}=\cos(\arg z_{kj}-\phi)$, *we then have*

$$\lim_{k\to\infty} g(t; W_k^{\phi}, c)=\left[\frac{2}{1+\sqrt{1+t^2}}\right]^c, \tag{4.1}$$

where g is defined by (3.1).

Proof. By Corollary 3.3, we have

$$g(t; W_k^{\phi}, c)=\prod_{j=1}^{k}(1-it\cos(\arg z_{kj}-\phi))^{-b_{kj}}.$$

Therefore,

$$\begin{aligned}\lim_{k\to\infty} g(t; W_k^{\phi}, c)&=\exp\left(-\lim_{k\to\infty}\sum_{j=1}^{k} b_{kj}\ln(1-it\cos(\arg z_{kj}-\phi))\right)\\&=\exp\left(-\frac{c}{2\pi}\int_0^{2\pi}\ln(1-it\cos(\theta-\phi))\,d\theta\right)\\&=\left[\exp\left(-(2\pi)^{-1}\int_0^{2\pi}\ln(1-it\cos\theta)\,d\theta\right)\right]^c.\end{aligned}$$

Since $\ln(1-it\cos\theta)=\ln|1-it\cos\theta|+i\arg(1-it\cos\theta)$ and $\arg(1-it\cos\theta)=\arg(1-it\cos(2\pi-\theta))$, we have

$$\begin{aligned}\int_0^{2\pi}\ln(1-it\cos\theta)\,d\theta&=\frac{1}{2}\int_0^{2\pi}\ln(1+t^2\cos^2\theta)\,d\theta\\&=2\pi\ln\frac{1+\sqrt{1+t^2}}{2}.\end{aligned}$$

The last identity is obtained by Formula 322.12b of Gröbner and Hofreiter (1973).

If we rotate axes counterclockwise through an angle ϕ, where $0\leqq\phi<\pi$, then z_{kj}^{ϕ} will be the projection of z_{kj} on the new first coordinate. Therefore, W_k^{ϕ} is the first coordinate with respect to rotated axes through an angle ϕ (counterclockwise) and

$W_k^0 = \mathbf{u}_{k*} \cdot \mathbf{x}_{k*}$. Notice that (4.1) holds for any real numbers c and the right-hand side of (4.1) does not depend on ϕ; that is, the limiting function of c-characteristic functions, $g(t; W_k^\phi, c)$, does not depend on ϕ, the angle corresponding to the transform of the marginal distribution W_k^0 under the regularity condition (A)-(D).

In the next lemma, we show that $\mathbf{V} = (X, Y)'$ has a spherical distribution if the first coordinate of any rotational transform of $\mathbf{V}$ has the same distribution as X. Therefore, under the regularity conditions (A)-(D), Lemma 4.1 together with the next lemma show that if $\mathbf{u}_{k*} \cdot \mathbf{z}_{k*}$ has a limiting distribution this limiting distribution is a spherical distribution.

LEMMA 4.2. *If the distribution of* $W^\phi = X \cos\phi + Y \sin\phi$, *where* $0 \leqq \phi < 2\pi$, *does not depend on* ϕ *(that is,* $W^{\phi_1} \sim W^{\phi_2}$ *for* $\phi_1 \neq \phi_2$*), then* $\mathbf{V} = (X, Y)'$ *has a spherical distribution.*

Proof. For any $\mathbf{t} = (t_1, t_2)' \neq \mathbf{0}$, there is one and only one pair (ρ, ϕ) such that $t_1 = \rho \cos\phi$, $t_2 = \rho \sin\phi$, $\rho > 0$ and $0 \leq \phi < 2\pi$. The characteristic function of $\mathbf{V} = (X, Y)'$ is

$$\begin{aligned} c(\mathbf{t}) &= E[\exp(i(t_1 X + t_2 Y))] \\ &= E[\exp(i\rho(X \cos\phi + Y \sin\phi))] \\ &= E[\exp(i\rho W^\phi)]. \end{aligned}$$

Since W^ϕ's have the same distribution for any ϕ, $c(\mathbf{t})$ depends on $\mathbf{t}$ only through ρ. But $\rho^2 = t_1^2 + t_2^2$, i.e., $\rho = (\mathbf{t}'\mathbf{t})^{1/2}$, and V has a spherical distribution if and only if $c(\mathbf{t})$ depends on $\mathbf{t}$ only through $\mathbf{t}'\mathbf{t}$. This completes the proof.

LEMMA 4.3. *Let* $\mathbf{V}_1 = (X_1, Y_1)'$ *and* $\mathbf{V}_2 = (X_2, Y_2)'$ *be two random vectors, whose first coordinates with respect to rotated axes through an angle* ϕ *are* W_1^ϕ *and* W_2^ϕ, *respectively, where* $0 \leqq \phi < 2\pi$. *We further assume that* $W_1^\phi \sim W_2^\phi$ *for any* ϕ. *Then* $\mathbf{V}_1 \sim \mathbf{V}_2$.

Proof. We denote the characteristic functions of $\mathbf{V}_1$ and $\mathbf{V}_2$ as $c_1(\mathbf{t})$ and $c_2(\mathbf{t})$, respectively. By the uniqueness of the characteristic function, we will complete the proof by showing that $c_1(\mathbf{t}) = c_2(\mathbf{t})$, for any real number pair $\mathbf{t}$. For any $\mathbf{t} = (t_1, t_2)' \neq 0$, there is one and only one pair (ρ, ϕ) such that $t_1 = \rho \cos\phi$, $t_2 = \rho \sin\phi$, $\rho > 0$ and $0 \leq \phi < 2\pi$. We have

$$c_j(\mathbf{t}) = E[\exp(i\rho W_j^\phi)] \quad \text{for } j = 1, 2.$$

But, since $W_1^\phi \sim W_2^\phi$, we therefore have $c_1(\mathbf{t}) = c_2(\mathbf{t})$.

LEMMA 4.4. *For any* $c > 0$, *let* $\tilde{Z} = (X, Y)'$ *be a vector random variable on the unit disk with its probability density function*

$$(c/\pi)(1 - x^2 - y^2)^{c-1}, \qquad 0 \leqq x^2 + y^2 < 1. \tag{4.2}$$

Furthermore, let $\tilde{Z}^\phi$ *be the first coordinate with respect to rotated axes through an angle* ϕ *(i.e.,* $\tilde{Z}^\phi = X \cos\phi + Y \sin\phi$*), where* $0 \leqq \phi < 2\pi$. *Then the c-characteristic function of* $\tilde{Z}^\phi$ *is*

$$g(t; \tilde{Z}^\phi, c) = \left[\frac{2}{1 + (1 + t^2)^{1/2}}\right]^c. \tag{4.3}$$

Proof. Without loss of generality, we need only to prove that (4.3) holds for $\phi = 0$, that is,

$$g(t; X, c) = \left[\frac{2}{1 + (1 + t^2)^{1/2}}\right]^c.$$

The probability density function of X is

$$2\int_0^{(1-x^2)^{1/2}}\left(\frac{c}{\pi}\right)(1-x^2-y^2)^{c-1}\,dy=\left(\frac{c}{\pi}\right)(1-x^2)^{c-1/2}\int_0^1 t^{-1/2}(1-t)^{c-1}\,dt$$

$$=(4\pi)(1-x^2)^{c-1/2}B\left(\frac{1}{2},c\right)$$

$$=(1-x^2)^{c-1/2}\Big/B\left(c+\frac{1}{2},\frac{1}{2}\right),\qquad -1<x<1.$$

The first identity follows by letting $y=t^{1/2}(1-x^2)^{1/2}$. The c-characteristic function of X is

$$g(t;X,c)=\int_{-1}^1\left\{(1-x^2)^{c-1/2}\Big/\left[(1-itx)^c\cdot B\left(c+\frac{1}{2},\frac{1}{2}\right)\right]\right\}dx$$

$$=R_{-c}\left(c+\frac{1}{2},c+\frac{1}{2};1+it,1-it\right)$$

$$=\left[\frac{2}{1+(1+t^2)^{1/2}}\right]^c.$$

The second and third identities can be obtained by using Exercises 5.1-3 and 6.10-12 in Carlson (1977). This completes the proof.

We have the following theorem on limiting distributions of linear combinations of Dirichlet vectors under the regularity conditions.

THEOREM 4.5. *Under the regularity conditions, the sequence of the random quantities* $\sum_{j=1}^k u_{kj}z_{kj}$, *where* $\mathbf{u}_{k*}\sim D(\mathbf{b}_{k*})$, *converges in distribution to a spherical distribution having probability density function* (4.2).

Proof. By Lemma 4.1, Lemma 4.4 and Theorem 3.5, the sequence of the distributions of the real part of $\sum_{j=1}^k u_{kj}z_{kj}$ converges in distribution to the marginal distribution of a distribution having probability density function (4.2). By Lemma 4.1 and Lemma 4.2, the limiting distribution has a spherical distribution. Lord (1954) shows that a spherical distribution is determined by its marginal distribution. This completes the proof.

An alternative proof is given directly by Lemma 4.1, Lemma 4.4, Theorem 3.5 and Lemma 4.3.

COROLLARY 4.6. *Assume the regularity conditions hold and denote* $V_c=\lim_{k\to\infty}\sum_{j=1}^k u_{kj}z_{kj}$, *where* $c=b_{k\cdot}$. *Then*

(a) V_1 *follows a uniform distribution on the unit disk with center* $(0,0)$.

(b) V_2 *follows the circularly symmetric distribution on the unit disk with probability density function*

$$f(x,y)=(2/\pi)(1-r^2),\qquad 0\leqq r<1,$$

where $r^2=x^2+y^2$.

(c) *Consider the limits in distributions,*

$$\lim_{c\to 0}V_c=U,\qquad \lim_{c\to\infty}V_c=V.$$

Then U has a uniform distribution on the unit circle and V has a point mass on the origin $(0,0)$.

COROLLARY 4.7. *If we have, for all k,*

$$b_{kj}=1/k,\qquad 1\leqq j\leqq k\qquad \text{so } b_{k\cdot}\equiv 1, \tag{4.4}$$

and $z_{k1}, z_{k2}, \cdots, z_{kk}$ are the kth roots of unity, then the limiting distribution of $\sum_{j=1}^{k} u_{kj} z_{kj}$, where $\mathbf{u}_{k*} \sim D(\mathbf{b}_{k*})$, will be a uniform distribution on the unit disk with its center at the origin.

Define the probability measure $\upsilon(v)$ on the unit disk by

$$d\upsilon(v) = (2/\pi)(1 - x_1^2 - x_2^2)\, dx_1\, dx_2$$

where $v = x_1 + ix_2$. By (2.7), Corollary 4.6(b) and the regularity conditions, we have the following theorem.

THEOREM 4.8. *The starlike function with representation* (2.1) *with uniform measure ν can be expressed by*

$$\zeta \cdot \int_{\Omega} (1 + v\zeta)^{-2}\, d\vartheta(v). \tag{4.5}$$

The more general starlike functions with representation (2.1) in terms of a more general measure ν would seem to be amenable to study through the Dirichlet random process, a random set function parameterized by a general probability measure (Ferguson (1973)).

Acknowledgments. I am deeply grateful to my adviser, James M. Dickey, for suggesting the problem and providing stimulating advice, encouragement, and helpful comments. I would also like to thank Louis Brickman and Benton Jamison for their help, and a referee for very helpful comments that led to considerable improvement of this paper.

REFERENCES

B. C. CARLSON (1977), *Special Functions of Applied Mathematics*, Academic Press, New York.

B. C. CARLSON AND D. B. SHAFFER (1984), *Starlike and prestarlike hypergeometric functions*, this Journal, 15, pp. 737–745.

K. L. CHUNG (1974), *A Course in Probability Theory*, Academic Press, New York.

L. DE BRANGES (1985), *A proof of the Bieberbach conjecture*, Acta Math., 154, pp. 1–2.

T. S. FERGUSON (1973), *A Bayesian analysis of some nonparametric problems*, Ann. Statist., 1, pp. 209–230.

W. GRÖBNER AND W. HOFREITER (1973), *Integraltafel*, Vol. 2, 5th ed., Springer-Verlag, Vienna, New York.

R. D. LORD (1954), *The use of the Hankel transformations in statistics. I. General theory and example*, Biometrika, 41, pp. 44–55.

G. SCHOBER (1975), *Univalent Functions—Selected Topics*, Springer-Verlag, New York.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

012

ON THE REDUCTION OF CONNECTION PROBLEMS FOR DIFFERENTIAL EQUATIONS WITH AN IRREGULAR SINGULAR POINT TO ONES WITH ONLY REGULAR SINGULARITIES, II*

W. BALSER†, W. B. JURKAT‡, AND D. A. LUTZ§

Abstract. Our purpose is to investigate lateral and central connection problems for systems of linear differential equations near an irregular singularity. In Part I [SIAM J. Math. Anal., 12 (1981), pp. 691–721] we showed how the lateral connection problem can be solved using some associated functions constructed from a formal fundamental solution. Here, we generalize the associated functions by introducing a complex parameter and show how certain values of these functions can be used to construct solutions in so-called Floquet form. We also show how the coefficients of the formal series can be asymptotically represented using other associated functions and how the central connection problems for the Floquet solution can be solved. We conclude with an application of the main results to the global solution of a rationalized form of Mathieu's equation that has two irregular singularities.

Key words. connection problems, irregular singularity, Floquet solutions

AMS(MOS) subject classifications. 34A20, 34C20

Introduction. We consider systems of linear differential equations of the form

$$x' = A(z)x = \left(\sum_0^\infty A_p z^{-p}\right)x, \tag{0.1}$$

where x is an n-dimensional column vector, the leading coefficient matrix A_0 has all distinct eigenvalues $\lambda_1, \cdots, \lambda_n$, and the power series converges for $|z| > a \geqq 0$. We will refer to these later on as our *basic assumptions.*

The classical theory of differential equations near an irregular singularity (see [4] or [18]) asserts that there exist *formal* fundamental solutions and two types of actual solutions that we term normal solutions and solutions in *Floquet form.* The global solution of (0.1) involves determining how the normal solutions are related to each other (lateral connection problem) and how the normal solutions are related to solutions in Floquet form (central connection problem). Our purpose here is to show how these connection problems can be solved with the aid of some *associated functions* that are constructed using the formal solutions.

Of these types of solutions, the formal ones are by far the easiest to construct (in general). There exist, as is well known, formal fundamental solutions of the form

$$H(z) = F(z)z^{\Lambda'} e^{\Lambda z}, \qquad F(z) = \sum_0^\infty F_p z^{-p}, \tag{0.2}$$

where $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}$, F_0 is any invertible matrix satisfying $F_0^{-1}A_0F_0 = \Lambda$, $\Lambda' = \operatorname{diag}\{F_0^{-1}A_1F_0\} \equiv \operatorname{diag}\{\lambda_1', \cdots, \lambda_n'\}$, the coefficients F_p are uniquely determined for $p \geqq 1$ (once F_0 is selected), and they can be calculated in a straightforward, recursive manner using only arithmetical operations. In what follows, we will assume that such a (fixed) formal fundamental solution has been constructed and we will base our subsequent discussion and calculations on it.

* Received by the editors July 8, 1985; accepted for publication (in revised form) March 16, 1987.

† Abteilung Mathematik V, Universität Ulm, 7900 Ulm/Donau, West Germany.

‡ Abteilung Mathematik V, Universität Ulm, 7900 Ulm/Donau, West Germany, and Department of Mathematics, Syracuse University, Syracuse, New York 13210.

§ Department of Mathematical Sciences, San Diego State University, San Diego, California 92182. The work of this author was supported by a grant from the National Science Foundation.

Next, using the classical asymptotic existence theorem, one can show that there is a family of actual solutions $\{X_\nu(z)\}$ (which we call normal solutions) and a covering of the Riemann surface of $\log z$, $|z|>a$, by sectorial regions $\{S_\nu\}$ (which are slightly larger than half-planes) such that $X_\nu(z)\cong H(z)$ as $z\to\infty$, $z\in S_\nu$, for every integer ν, and the $X_\nu(z)$ are even uniquely determined (individually) by this condition. (In Part I of this paper, reference [1], we showed how normal solutions can be constructed either using Laplace transforms of the associated functions or by summing the formal series as a factorial series.)

Finally, a well-known classical result (analogous to Floquet's theorem for differential equations with periodic coefficients) asserts that every fundamental solution $X(z)$ may be expressed in the form

$$X(z)=L(z)z^M, \tag{0.3}$$

where $L(z)=\sum_{-\infty}^{+\infty} L_p z^p$ is single-valued and analytic in $a<|z|<\infty$, and the constant matrix M is called a *monodromy matrix.* We say that such a solution is expressed in *Floquet form.*

In light of these facts, it is natural to consider the following *connection problems,* which explain the global behavior of solutions:

(i) Given two consecutive normal solutions, determine the constant matrices V_ν (called the *Stokes' multipliers*) that satisfy

$$X_{\nu-1}(z)=X_\nu(z)V_\nu.$$

(ii) Given a solution $X(z)$ in Floquet form, determine its asymptotic as $z\to\infty$, $z\in S_\nu$, i.e., find the *central connection factors* Ω_ν defined by

$$X(z)=X_\nu(z)\Omega_\nu.$$

(iii) Given a fundamental solution matrix with a known asymptotic near ∞, say one of the normal solutions, express it in Floquet form, i.e., determine a monodromy matrix and the corresponding Laurent coefficients.

Using the fact that the normal solutions (and hence the Stokes' multipliers are uniquely determined by the selected $H(z)$, one can conclude that there should exist relations (at least in an abstract sense) between the quantities in $H(z)$ and the Stokes' multipliers. Likewise, since a selected actual solution (say in Floquet form) and $H(z)$ uniquely determine the central connection factors, there ought to exist relations between these quantities as well. Our objective is to construct some concrete relations of these types; more specifically, our main results concern representation formulas for the formal and Laurent coefficients, which can be applied to solve the connection problems.

The derivation of these results depends upon developing and applying properties of some generalizations of the associated functions we have considered in [1]. They not only interpolate the coefficients in the formal series, but are also a natural and convenient way of assembling the information present in the formal solution that is relevant for our problems. The particular associated functions we construct now also involve a complex parameter s, corresponding to the "shifted" equations

$$x_s'=(A(z)-sz^{-1}I)x_s$$

that are obtained from (0.1) by the transformation $x=x_s z^s$. In addition to studying their analytic properties and connection phenomena in the variable t, we also require their analytic and asymptotic properties with respect to s.

Aside from this direct role that the parameter plays in our results, it also has a more indirect, unifying role through the discussion of a nonstandard type of difference

equation which the associated functions satisfy and whose appearance we motivate as follows.

For the often-studied case of the "two-term" differential equation

$$x' = (A_0 + A_1 z^{-1})x, \tag{0.4}$$

one constructs the corresponding system of linear difference equations (in the variable s and parameter t)

$$s\xi(s+1, t) + t\xi(s, t) = A_0\xi(s, t) + A_1\xi(s+1, t), \tag{0.5}$$

which has central importance because both the columns of the coefficients F_p in the formal series as well as the columns of the Laurent coefficients L_p (in case A_1 has incongruent (mod 1) eigenvalues) are special solutions for appropriate values of s and t. For example, for a suitable solution ξ of (0.5), $\xi(\lambda_k' - p + 1, \lambda_k) = f_k(p)$, the kth column of F_p, while (for a generally different ξ) $\xi(\mu_j + p + 1, 0) = l_j(p)$, the jth column of L_p (when $\mu_1, \cdots, \mu_n$ are the eigenvalue of A_1 and $M = \operatorname{diag}\{\mu_1, \cdots, \mu_n\}$). Using this correspondence, Okubo [15] and Kohno [14] have derived relationships between the coefficients of the formal series, the Laurent coefficients, and the central connection factors. (See also Hopf [8] and Knobloch [12].) R. Schäfke [17] has also considered connection problems for (0.4) and has derived relations which involve both the lateral and central connection problems. Some of our results may be thought of as extensions of these to the "general" case (0.1). Our methods are closer to Schäfke's than Okubo and Kohno's, which rely much more heavily on the interplay with the difference equation.

For a "general" equation (0.1), one can also analogously consider the "difference equation" (in the variable s and parameter t)

$$s\xi(s+1, t) + t\xi(s, t) = \sum_0^\infty A_\nu \xi(s+\nu, t). \tag{0.6}$$

Even in this case it is easy to see that both $f_k(p)$ and $l_j(p)$ are (at least in a formal sense) solutions of (0.6). While the vectors $f_k(p)$ can still be recursively calculated from (0.6) (note that the series on the right-hand side terminates because $f_k(p) = 0$ for $p < 0$), the same is not true for the coefficients $l_j(p)$ of the Laurent series. This happens because, in general, the Laurent series is a doubly infinite series and one is led to an infinite system of coupled linear equations for which no procedure to recursively calculate the terms is known. Von Koch [13] has applied the theory of infinite determinants to the study of these equations, leading to an approximation of a Floquet-type solution, but his method appears to differ substantially from the approach we take to construct Floquet solutions. (On the other hand, our approach is restricted to equations with a pole type singularity, while the procedure of von Koch applies, in principle, to cases where $A(z) = \sum_{-\infty}^{\infty} A_p z^{-p}$.)

Analogous to the case of (0.5), we call (0.6) a *difference equation of modified first order.* In contrast to (0.5) where the asymptotic theory of linear difference equations applies, it is somewhat surprising that certain explicit solutions of (0.6) can be constructed that have a known asymptotic, both as $\operatorname{Re} s \to +\infty$ (Proposition 4) and $\operatorname{Re} s \to -\infty$ (Proposition 5). In fact, the connection problems we consider can all be rephrased in terms of connection problems for particular solutions of (0.6). But the difference equation plays a more indirect role in our development because it may have many other solutions than the ones we consider, which seem to have no relevance to the differential equation. One reason behind the fact that we can construct particular solutions to (0.6) and discuss their complete analytic and asymptotic behavior is that

there exists a Birkhoff reduction (see, e.g., [1] for the definition) taking (0.1) into some equation of the form (0.4) (but not with the same A_0 and A_1). The Birkhoff transformation can be carried over to relate certain solutions of (0.6) and corresponding ones of (0.5). But solutions of (0.6) that grow too rapidly (as $\operatorname{Re} s \to +\infty$) cannot be thus transformed and one can see that only a certain n-dimensional subspace of solutions of (0.6) can actually correspond to solutions of (0.5).

As a final remark, we wish to alert the reader to the following situation: The values of the analytic continuations of the associated functions and the constants entering their connection relations play a key role in our calculations. Since the associated functions have, in general, branch points at $\lambda_1, \cdots, \lambda_n$, it is necessary before making their continuations to cut the plane near these points and specify a value for $\arg(t-\lambda_k)$ for t close to λ_k. The values are then strongly dependent on which system of cuts is used. For each particular application, we choose a system of cuts that most easily allows the quantities entering the relations to be identified as, for example, certain Stokes' multipliers or central connection factors. Thus in §§ 2–5 we use one type of cuts for the lateral connection problem, whereas in §§ 6 and 7 we use another type for the central connection problem. Both of these differ from the systems of cuts we used in [1], which were particularly convenient for integral representations.

1. Some relations between the connection matrices. Throughout this paper we consider a fixed, but arbitrary, differential equation (0.1) satisfying our basic assumptions in the Introduction. We also make a fixed, but arbitrary, selection of an enumeration of the eigenvalues $\lambda_1, \cdots, \lambda_n$ of A_0 and consider any fixed formal fundamental solution matrix $H(z)$ of the form (0.2). Such an $H(z)$ always exists and is unique up to a constant invertible diagonal right-hand factor D. The freedom in the choice of $H(z)$ corresponds exactly to the choice of F_0.

Letting $\{X_\nu(z)\}$ denote the normal solutions corresponding to $H(z)$ and (V_ν), the associated Stokes' multipliers, one sees that the *circuit factor* $e^{2\pi i M_\nu}$ for $X_\nu(z)$ (defined by $X_\nu(z\,e^{2\pi i}) = X_\nu(z)\,e^{2\pi i M_\nu}$) is given by

$$e^{2\pi i M_\nu} = e^{2\pi i \Lambda'} V_{\nu+m} \cdots V_{\nu+1} \tag{1.1}$$

(compare [3, Part II, Prop. 4]).

Remark 1.1. If we fix an arbitrary system $\mathcal{R}$ of representatives modulo one for the complex numbers, then there always exists a matrix M_ν having its eigenvalues in $\mathcal{R}$, such that (1.1) holds, and in fact M_ν is unique (to see the uniqueness, one can use a proposition in [10, p. 38]). From $V_{\nu+m} = e^{-2\pi i \Lambda'} V_\nu\, e^{2\pi i \Lambda'}$ [10] we conclude

$$e^{2\pi i M_\nu} = V_\nu\, e^{2\pi i M_{\nu-1}} V_\nu^{-1},$$

i.e., a possible choice for $M_{\nu-1}$ having eigenvalues in $\mathcal{R}$ is $V_\nu^{-1} M_\nu V_\nu$, and from the uniqueness stated above we conclude that

$$M_{\nu-1} = V_\nu^{-1} M_\nu V_\nu \quad \text{for every } \nu. \tag{1.2}$$

According to (1.2), the Jordan canonical form of M_ν is independent of ν and may be denoted by M. Then M is unique, if we choose it to be upper triangular and assume its blocks ordered according to some arbitrarily fixed rule (see [10], p. 34). If we replace $H(z)$ by $H(z)\,D$ (with D as above), then V_ν resp. M_ν are replaced by $D^{-1} V_\nu D$ resp. $D^{-1} M_\nu D$; hence, M is even independent of the choice of $H(z)$ and therefore *corresponds uniquely to the differential equation* (0.1).

According to the above construction, there exists a fundamental solution matrix $X(z)$ for (0.1) having M as a monodromy matrix, i.e.,

$$L(z) = X(z) z^{-M} \tag{1.3}$$

is single-valued and analytic in $|z|>a$. Such a fundamental solution is determined up to a constant, invertible, right-hand factor C that commutes with M (using again the proposition in [10, p. 38] or Lemma 1 of [3, Part I]). For each such $X(z)$, there exists (for every integer ν) a constant invertible matrix Ω_ν called the νth central connection factor of $X(z)$ such that

$$X(z)=X_\nu(z)\Omega_\nu, \qquad |z|>a. \tag{1.4}$$

(If $X(z)$ is replaced by $X(z)C$ with C as above, the Ω_ν is to be replaced by $\Omega_\nu C$.)

Remark 1.2. It follows immediately from the preceding discussion that the knowledge of the matrices

$$e^{2\pi i\Lambda'}, V_{\nu+1}, \cdots, V_{\nu+m}$$

for any fixed, but arbitrary, ν determines a solution $X(z)$ in Floquet form and its central connection factors up to within their natural degree of freedom. To see this, let M_ν denote the unique matrix satisfying (1.1) (with its eigenvalues coming from $\mathcal{R}$). If ν_0 is an arbitrarily fixed integer and Ω_{ν_0} is any invertible matrix satisfying

$$M=\Omega_{\nu_0}^{-1}M_{\nu_0}\Omega_{\nu_0},$$

define

$$X(z)=X_{\nu_0}(z)\Omega_{\nu_0}, \qquad |z|>a.$$

Then (1.4) determines the matrices Ω_ν for every integer ν, and it is clear from the definition of V_ν that

$$\Omega_\nu = V_\nu\Omega_{\nu-1} \tag{1.5}$$

for every integer ν. From the definition of M we find that $L(z)=X(z)z^{-M}$ is single-valued and analytic in $|z|>a$; hence, it has a Laurent expansion. We shall show later (in §§ 6 and 7) how the Laurent coefficients may be calculated using generalizations of the associated functions considered in [1].

Since M is selected in some particular Jordan form with its eigenvalues from a fixed system of representatives modulo one, then Ω_ν satisfying

$$e^{2\pi iM}=\Omega_\nu^{-1}\,e^{2\pi iM_\nu}\Omega_\nu$$

is determined up to a right-hand invertible factor that commutes with $e^{2\pi iM}$ (i.e. with M). This corresponds exactly to the freedom in the choice of a fundamental solution $X(z)$ for which $X(z)z^{-M}$ is single-valued.

It is clear from (1.4) and the definition of V_ν that the knowledge of two consecutive central connection factors determines a Stokes' multiplier. We now show that any *one* central connection factor Ω_ν together with the knowledge of the circuit factor $e^{2\pi iM}$ determine *all* the Stokes' multipliers and consequently *all* the other Ω_l.

PROPOSITION 1. *Consider a fixed, but arbitrary, differential equation* (0.1) *satisfying our basic assumptions and any selected formal fundamental solution matrix* $H(z)$ *as above. Let* $X(z)$ *denote an arbitrary fundamental solution matrix of* (0.1) *and assume that the circuit factor* $e^{2\pi iM}$ (*with* M *not necessarily in canonical form*) *is known along with one central connection factor* Ω_ν *for any fixed, but arbitrary integer* ν. *Then the formal circuit factor* $e^{2\pi i\Lambda'}$ *and all the normalized Stokes' multipliers* V_l, $-\infty<l<+\infty$, *can be explicitly computed*; *hence, all the remaining central connection factors* Ω_l *can be explicitly computed* (*without using the differential equation*).

Proof. Using (1.4) with $z=z_0$ and $z=z_0\,e^{2\pi i}$ we obtain

$$\Omega_\nu\,e^{2\pi iM}\Omega_\nu^{-1}=e^{2\pi iM_\nu}=e^{2\pi i\Lambda'}V_{\nu+m}\cdots V_{\nu+1}. \tag{1.6}$$

For convenience in this argument, we may assume that $\lambda_1, \cdots, \lambda_n$ are enumerated so that (with $\mu = m/2$) the matrices $V_{\nu+m}, \cdots, V_{\nu+\mu+1}$ are all upper triangular while $V_{\nu+\mu}, \cdots, V_{\nu+1}$ are all lower triangular (cf. [1, § 3]). Then

$$U = V_{\nu+m} \cdots V_{\nu+\mu+1}, \quad \text{resp. } L = V_{\nu+\mu} \cdots V_{\nu+1},$$

are also upper, resp. lower, triangular matrices and both have all ones on the diagonal (since all the V_l have ones on the diagonal), hence (1.6) shows that $\Omega_\nu e^{2\pi i M} \Omega_\nu^{-1}$ can be factored as $e^{2\pi i \Lambda'} UL$. Whenever such a factorization is possible, it can easily be shown that the factors are unique since the diagonals of U and L are prescribed. Moreover, the factors $e^{2\pi i \Lambda'}$, U and L can be explicitly calculated using only arithmetical operations (see [6, p. 33] for the explicit formulas). From U and L, all the components $V_{\nu+m}, \cdots, V_{\nu+1}$ can be uniquely calculated (see [10, p. 80]) and using

$$V_{l+m} = e^{-2\pi i \Lambda'} V_l e^{2\pi i \Lambda'} \tag{1.7}$$

for every integer l, we obtain the complete collection of Stokes' multipliers and consequently all the other Ω_l from (1.5).

Remark 1.3. As mentioned in the Introduction, for a differential equation (0.4) one can always calculate a Floquet solution in a straightforward manner since 0 is a singularity of the first kind. In such a case the central connection problem is reduced to the calculation of a corresponding matrix Ω_ν for any fixed integer ν. Proposition 1 then tells us that all the lateral connection matrices and all the other central connection matrices can be explicitly calculated with arithmetical operations.

Suppose that a fixed solution $X(z) = L(z) z^M$ in Floquet form is known, say with M in a particular Jordan canonical form and whose eigenvalues come from a fixed system of representatives. If we also know one of the matrices M_ν (for some fixed integer ν), then (1.6) was seen to determine Ω_ν up to a right-hand invertible factor that commutes with M, and the computation of Ω_ν modulo this freedom is an algebraic problem. Moreover, as a consequence of Abel's formula

$$z^{\operatorname{tr} M} \det L(z) = c\, e^{z \operatorname{tr} A_0} z^{\operatorname{tr} \Lambda'} \exp\left[-\sum_2^\infty \operatorname{tr} A_\nu \frac{z^{-\nu+1}}{\nu - 1} \right], \tag{1.8}$$

(with a nonzero constant c) we now show that $\det \Omega_\nu$ is also determined.

PROPOSITION 2. *Consider a fixed, but arbitrary, differential equation* (0.1), *a selected formal fundamental solution* $H(z)$ *as above, and a given solution* $X(z) = L(z) z^M$. *If* Ω_ν *satisfies* (1.4), *then*

$$z^{\operatorname{tr} M} \det L(z) = \det F_0 \det \Omega_\nu z^{\operatorname{tr} \Lambda'} e^{z \operatorname{tr} A_0} \exp\left[-\sum_2^\infty \operatorname{tr} A_\nu \frac{z^{-\nu+1}}{\nu - 1} \right].$$

Proof. From (0.3) we have

$$\det X(z) = \det X_\nu(z) \det \Omega_\nu, \quad z \in S_\nu,$$

while from Abel's formula (1.8) we find (for a suitable constant $c_\nu \neq 0$)

$$\det X_\nu(z) = c_\nu\, e^{z \operatorname{tr} A_0} z^{\operatorname{tr} \Lambda'} \exp\left[-\sum_2^\infty \operatorname{tr} A_\nu \frac{z^{-\nu+1}}{\nu^{-1}} \right]. \tag{1.9}$$

Hence $c = c_\nu \det \Omega_\nu$. To evaluate the constants, first note that

$$\det X_\nu(z) \cong z^{\operatorname{tr} \Lambda'} e^{\operatorname{tr} \Lambda z} \det F(z) \quad \text{as } z \to \infty \text{ in } S_\nu;$$

hence observing $\operatorname{tr} A_0 = \operatorname{tr} \Lambda$ we obtain

$$c_\nu = \det F_0 \quad \text{for all } \nu.$$

This completes the proof.

Remark 1.4. In the special case that $A(z) = \Lambda + A_1 z^{-1}$ and A_1 has incongruent (but possibly equal) eigenvalues μ_i, $1 \leqq i \leqq n$ (modulo one), then M may either be selected to be equal to A_1 or to a particular Jordan canonical form of A_1. In both cases

$$L(z) = \sum_0^\infty L_p z^p$$

with L_0 invertible. In case $M = A_1$ a natural selection for L_0 is I while if M is a particular Jordan canonical form we could select L_0 to consist of certain eigenvectors and generalized eigenvectors of A_1 having determinant also equal to one. Then using

$$\operatorname{tr} M = \sum_1^n \mu_i = \sum_1^n \lambda_j'$$

and letting $z \to 0$ as in the statement of Proposition 2 we obtain

$$1 = \det L_0 = \det F_0 \det \Omega_\nu \quad \text{for all } \nu.$$

If we also select $\det F_0 = 1$, then

$$\det \Omega_\nu = 1 \quad \text{for all } \nu.$$

From the fact that the matrices $e^{2\pi i M}$ and $e^{2\pi i \Lambda'} V_{\nu+m} \cdots V_{\nu+1}$ are similar, one can derive some other interesting relations which explain in a certain quantitative sense how the presence of nontrivial Stokes' multipliers accounts for the difference between the *formal circuit factor* $e^{2\pi i \Lambda'}$ and an *actual circuit factor* $e^{2\pi i M}$. To see this, observe that because of the similarity,

$$\sigma_k(e^{2\pi i M}) = \sigma_k(e^{2\pi i \Lambda'} V_{\nu+m} \cdots V_{\nu+1}), \tag{1.10}$$

where σ_k denotes the kth symmetric function of the matrix (i.e. $\pm$ kth coefficient in the characteristic polynomial). For $k = n$ this gives the obvious relation $\det(e^{2\pi i M}) = \det(e^{2\pi i \Lambda'} V_{\nu+m} \cdots V_{\nu+1})$, which implies

$$\sum_1^n \mu_i \equiv \sum_1^n \lambda_j' \pmod 1,$$

but for $1 \leqq k \leqq n-1$ the equations contain some deeper and more interesting information. For example, in the case of a "two-term" differential equation (0.4), recall that if A_1 has all incongruent (but possibly equal) eigenvalues (modulo one), then a choice for M is A_1 itself. Hence (1.10) can be written as

$$\sigma_k(e^{2\pi i A_1}) = \sigma_k(e^{2\pi i \Lambda'} V_{\nu+m} \cdots V_{\nu+1}), \qquad k = 1, 2, \cdots, n. \tag{1.11}$$

Even if A_1 has congruent eigenvalues modulo one, we claim that (1.11) still holds by analytic perturbation because the quantities on both sides of the equation are analytic functions of the entries of A_1 (when A_0, F_0 and the diagonal Λ' are fixed and the off-diagonal elements of A_1 are allowed to vary) and in this way one can always arrange for A_1 to have incongruent eigenvalues.

In the special case of (1.10) when $n=2$ and $k=1$, one can rewrite the equation by introducing an auxiliary parameter

$$\mu=\mu_1-\frac{\lambda_1'+\lambda_2'}{2}, \qquad -\mu=\mu_2-\frac{\lambda_1'+\lambda_2'}{2}$$

(where we use that $\mu_1+\mu_2=\lambda_1'+\lambda_2'$), and letting

$$V_2=\begin{bmatrix}1 & c_2\\ 0 & 1\end{bmatrix}, \qquad V_1=\begin{bmatrix}1 & 0\\ c_1 & 1\end{bmatrix}$$

(here we assume that λ_1 and λ_2 are ordered lexicographically). Then a short calculation shows that

$$2\cos 2\pi\mu=2\cos\pi(\lambda_2'-\lambda_1')+c_1c_2\,e^{\pi i(\lambda_1'-\lambda_2')}, \tag{1.12}$$

and using the explicit expressions for c_1, c_2 in terms of some invariants γ, γ' that are related to the asymptotic of the coefficients in the formal series (see [10, p. 182]), we see that (1.12) can be expressed as

$$\cos 2\pi\mu=\cos\pi(\lambda_2'-\lambda_2')-2\pi^2\gamma\gamma'. \tag{1.13}$$

This equation, which played a key role in the inverse (matching) problem of constructing a two-term equation with prescribed invariants (see [11, § 7]), was called for obvious reasons the "cosine equation." Thus the equations (1.11) are natural generalizations of this equation for larger-dimensional systems of differential equations.

2. Asymptotic behavior of the coefficients in a formal solution. In this section we consider a fixed, but arbitrary, differential equation (0.1) that satisfies our basic assumptions and a fixed, but arbitrary, formal fundamental solution matrix $H(z)$ of the form (0.2). We also assume that a scalar shift $x=z^\gamma\hat{x}$ has been made with an appropriate (real or complex) number γ so that

$$\Lambda'=\operatorname{diag}\{F_0^{-1}A_1F_0\}=\operatorname{diag}\{\lambda_1',\cdots,\lambda_n'\}$$

has the property that (for each $j=1,2,\cdots,n$) λ_j' is not an integer and $\operatorname{Re}\lambda_j'<0$.

Recall from [1, p. 693] that for each integer k, $1\leqq k\leqq n$, associated functions $y_k(t)$ were constructed as

$$y_k(t)=\sum_{p=0}^{\infty} f_k(p)\Gamma(\lambda_k'+1-p)(t-\lambda_k)^{p-\lambda_k'-1},$$

where $f_k(p)$ denotes the kth column of F_p and the series converges for $|t-\lambda_k|$ sufficiently small. Since for each natural number p the function

$$(t-\lambda_k)^{\lambda_k'-p}y_k(t)$$

is single-valued and analytic in a small deleted neighborhood of λ_k, we obtain (integrating in the positive direction and for $\varepsilon>0$ sufficiently small) from Cauchy's theorem

$$f_k(p)\Gamma(\lambda_k'+1-p)=\frac{1}{2\pi i}\oint_{|t-\lambda_k|=\varepsilon}(t-\lambda_k)^{\lambda_k'-p}y_k(t)\,dt. \tag{2.1}$$

To determine the explicit asymptotic behavior of the integral as $p\to+\infty$, we require knowledge of the function $y_k(t)$ away from the point λ_k. Since (see [1, § 4.2]) $y_k(t)$ is analytic everywhere in the finite complex t-plane, except for possible branch points at λ_j, $1\leqq j\leqq n$, in order to have a definite means for the analytic continuation of $y_k(t)$ outside of $|t-\lambda_k|<\varepsilon$ we will cut the plane at each of these points along rays in certain

directions extending to ∞. Such a direction η is called admissible (with respect to λ_j) if the cut $\arg(t-\lambda_j)=\eta$ contains no point λ_l, $l\neq j$. It is also natural to make the cuts nonintersecting so that all points are accessible without having to cross any of the cuts.

One method of making cuts that was convenient for the purposes in [1] is to make all the cuts have the same direction; we spoke of these as "parallel cuts." For our present purposes it is natural to make the following "star-shaped" systems of cuts:

Let k be a fixed, but arbitrary integer, $1\leqq k\leqq n$, and let η be any fixed, admissible direction (with respect to λ_k). Then we make a cut from λ_k to ∞ along $\arg(t-\lambda_k)=\eta$, and from each point λ_j we cut along

$$\arg(t-\lambda_j)=\eta_{jk}=\arg(\lambda_j-\lambda_k)\in(\eta-2\pi,\eta),\qquad j\neq k,\quad 1\leqq j\leqq n,$$

if η_{jk} is admissible with respect to λ_j, i.e., if no other λ_l lies on $\arg(t-\lambda_j)=\eta_{jk}$; otherwise, turn the cut slightly to the right so that it becomes admissible with respect to λ_j. Along the right border of each cut (looking toward ∞) we select

$$\arg(t-\lambda_j)=\eta_{jk}-0,\quad \text{resp., } \arg(t-\lambda_k)=\eta-0$$

and we use this choice of the argument in identifying the branch of $\log(t-\lambda_j)$, resp., $\log(t-\lambda_k)$, to be used in defining nonintegral powers of $(t-\lambda_j)$, resp., $(t-\lambda_k)$. We denote the complex t-plane with this system of cuts and choices of the arguments by $\mathcal{P}_{k,\eta}$.

It is easy to check that the proof of Theorem 1 in [1, § 4.2] extends to $\mathcal{P}_{k,\eta}$; hence, the associated function $y_k(t)=y_k(t;\eta)$ can be analytically continued to all of $\mathcal{P}_{k,\eta}$ and there exist constants $c_{jk}=c_{jk}(\eta)$ such that

$$y_k(t)=c_{jk}y_j(t)+\operatorname{reg}(t-\lambda_j),\qquad t\in\mathcal{P}_{k,\eta},\tag{2.2}$$

where $y_j(t)=y_j(t;\eta_{jk})$ is the analytic continuation (in $\mathcal{P}_{k,\eta}$) of the function defined near λ_j by the convergent series with the meaning for the nonintegral power as above. Note that these functions and constants depend upon the choice of the cuts, but to simplify the notation we do not specifically display that dependence. (These functions and constants also generally differ from the analogous quantities treated in [1] in the case of parallel cuts but for which we use the same notation.)

In order to show how the constants c_{jk} are related to certain Stokes' multipliers, we recall from Theorem 2 of [1, pp. 714–715] that for any fixed integer ν and any α between the critical rays $\eta_{\nu+1}$ and η_ν (see [1, § 3.1] for the definition of critical rays) we have

$$x_{\nu,j}(z)=\frac{1}{2\pi i}\int_{\gamma_j(\alpha)}e^{zt}y_j(t)\,dt,\qquad z\in\mathcal{S}(\alpha),\quad 1\leqq j\leqq n,$$

where

$$\mathcal{S}(\alpha)=\left\{\operatorname{Re}(z\,e^{i\alpha})<-a;\ \frac{\pi}{2}-\alpha<\arg z<\frac{3\pi}{2}-\alpha\right\},$$

and $\gamma_j(\alpha)$ denotes the loop encircling the ray $\arg(t-\lambda_j)=\alpha$ in the positive direction and such that no λ_l, $l\neq j$, is contained on or inside this loop. We also take $\arg(t-\lambda_j)\in(\alpha-2\pi,\alpha)$ for t on this loop. Since we are assuming $\operatorname{Re}\lambda_j'<0$, we may deform this loop integral into a ray integral as

$$x_{\nu,j}(z)=\frac{1-e^{2\pi i\lambda_j'}}{2\pi i}\int_{\lambda_j}^{\infty(\alpha)}e^{zt}y_j(t)\,dt,\qquad z\in\mathcal{S}(\alpha),\quad 1\leqq j\leqq n,\tag{2.3}$$

if we integrate along $\arg(t-\lambda_j)=\alpha$ and define $y_j(t)$ consistent with this selection of $\arg(t-\lambda_j)$.

Now let ν denote any fixed, but arbitrary, integer satisfying $\eta - 2\pi < \eta_\nu < \eta$ and let α, $\tilde{\alpha}$ be selected close to each other and satisfy

$$\eta_{\nu+1} < \alpha < \eta_\nu < \tilde{\alpha} < \eta_{\nu-1}.$$

Then from (2.3) we have for $z \in \mathcal{S}(\alpha) \cap \mathcal{S}(\tilde{\alpha}) \subset S(\tau_\nu - \pi, \tau_\nu)$

$$x_{\nu-1,k}(z) - x_{\nu,k}(z) = \frac{1 - e^{2\pi i \lambda_k'}}{2\pi i} \left\{ \int_{\lambda_k}^{\infty(\alpha)} - \int_{\lambda_k}^{\infty(\tilde{\alpha})} \right\} e^{zt} y_k(t)\, dt$$

and the selection of the branch of $y_k(t)$ in the integrals coincides with the selection in $\mathcal{P}_{k,\eta}$, provided we take α and $\tilde{\alpha}$ sufficiently close to η_ν (note that if for several $j \neq k$ we have $\arg(\lambda_j - \lambda_k) = \eta_\nu$, then without loss of generality we may assume that the cuts from those λ_j to ∞ do not intersect with the ray $\arg(t - \lambda_k) = \alpha$). It is easily seen from the definition of critical rays that those indices $j \neq k$ satisfying $\eta_\nu = \eta_{jk}$ are precisely the ones for which λ_j lies on $\arg(t - \lambda_k) = \eta_{jk}$. From the definition of the position set ρ_ν (see [1, § 3.1]) we conclude that this happens iff $(j, k) \in \rho_\nu$. For all these j the cuts from λ_j to ∞ have the same direction $\eta_\nu - \varphi = \alpha$, and with the help of Cauchy's theorem we conclude (using (2.2)) (for k as above)

$$\begin{aligned} x_{\nu-1,k}(z) - x_{\nu,k}(z) &= \frac{e^{2\pi i \lambda_k'} - 1}{2\pi i} \sum_{\substack{j \\ (j,k) \in \rho_\nu}} \int_{\gamma_j(\alpha)} e^{zt} y_k(t)\, dt \\ &= (e^{2\pi i \lambda_k'} - 1) \sum_{\substack{j \\ (j,k) \in \rho_\nu}} c_{jk} \frac{1}{2\pi i} \int_{\gamma_j(\alpha)} e^{zt} y_j(t)\, dt \\ &= (e^{2\pi i \lambda_k'} - 1) \sum_{\substack{j \\ (j,k) \in \rho_\nu}} c_{jk} x_{\nu,j}(z). \end{aligned}$$

Comparing this to $X_{\nu-1}(z) = X_\nu(z) V_\nu$, we find that for every j with $(j, k) \in \rho_\nu$ the constant $(e^{2\pi i \lambda_k'} - 1) c_{jk}$ coincides with the element of V_ν in the (j, k) position. We denote this element by v_{jk} (note that in a sense, this does not depend upon ν as long as we restrict to such ν with $\eta - 2\pi < \eta_\nu < \eta$, since then to every $j \neq k$ there exists exactly one such ν for which $(j, k) \in \rho_\nu$), and we then have for every fixed k, $1 \leqq k \leqq n$,

$$c_{jk} = (e^{2\pi i \lambda_k'} - 1)^{-1} v_{jk}, \qquad j \neq k, \quad 1 \leqq j \leqq n. \tag{2.4}$$

Now consider loops in $\mathcal{P}_{k,\eta}$ which we denote by $l_j(R)$ $(1 \leqq j \leqq n, j \neq k)$, and which have the following properties.

The loop $l_j(R)$ (for sufficiently large $R > 0$) extends from the point on the left border of the cut from λ_j to ∞ (looking towards ∞) for which $|t - \lambda_k| = R$ to the corresponding point on the right border, encircling the point λ_j in the positive direction and no other one of the points $\lambda_1, \cdots, \lambda_n$.

Then deforming the contour in (2.1), using (2.2), (2.4) and observing that

$$(t - \lambda_k)^{\lambda_k' - p} \operatorname{reg}(t - \lambda_j)$$

is holomorphic on $l_j(R)$ and its interior, we obtain

$$f_k(p) = \sum_{j \neq k} v_{jk} \varphi_{kj}(R, p) + I_k(R, p), \tag{2.5}$$

where

$$\begin{aligned} \varphi_{kj}(R, p) &= \{2\pi i \Gamma(\lambda_k' + 1 - p)(1 - e^{2\pi i \lambda_k'})\}^{-1} \int_{l_j(R)} (t - \lambda_k)^{\lambda_k' - p} y_j(t)\, dt \\ &= \frac{\Gamma(p - \lambda_k')}{(2\pi i)^2} e^{\pi i (p - \lambda_k')} \int_{l_j(R)} (t - \lambda_k)^{\lambda_k' - p} y_j(t)\, dt, \end{aligned} \tag{2.6}$$

and

$$I_k(R,p)=(2\pi i\Gamma(\lambda_k'+1-p))^{-1}\int_{|t-\lambda_k|=R}(t-\lambda_k)^{\lambda_k'-p}y_k(t)\,dt. \tag{2.7}$$

Note that in the integral representation of $I_k(R,p)$ the integration actually takes place over a set of $(n-1)$ disjoint segments on $|t-\lambda_k|=R$, and since $y_k(t)$ is bounded on each such segment, it is easy to see that for R as above (fixed) and every $\varepsilon>0$ (small)

$$I_k(R,p)=\Gamma(p)O((R-\varepsilon)^{-p})\quad \text{as } p\to\infty. \tag{2.8}$$

Since $\operatorname{Re}\lambda_j'<0$, the function $y_j(t)$ is integrable at λ_j, and deforming the contour of integration in (2.6) yields

$$\varphi_{kj}(R,p)=(1-e^{2\pi i\lambda_j'})\frac{\Gamma(p-\lambda_k')}{(2\pi i)^2}\,e^{\pi i(p-\lambda_k')}\int_{\lambda_j}^{\lambda_j+t_{jk}}(t-\lambda_k)^{\lambda_k'-p}y_j(t)\,dt,$$

where the integration takes place along the *right border* of the cut and accordingly $|t_{jk}|=R$, $\arg t_{jk}=\eta_{jk}-\varepsilon_{jk}$ with a suitably chosen small $\varepsilon_{jk}>0$. To find the asymptotic behavior of $\varphi_{jk}(R,p)$ as $p\to\infty$, we may (with the help of Cauchy's theorem) replace the integral from λ_j to λ_j+t_{jk} by a sum of two integrals from λ_j to $\lambda_j+\rho\exp[i\eta_{jk}]$ resp. from $\lambda_j+\rho\exp[i\eta_{jk}]$ to λ_j+t_{jk} (with sufficiently small $\rho>0$). Since ε_{jk} was taken small, the second integral may be easily estimated as $O((|\lambda_j-\lambda_k|+\rho/2)^{-p})$ as $p\to\infty$. In the first integral, we make the change of variable

$$t=t(s)=\lambda_j+(e^s-1)(\lambda_j-\lambda_k),\qquad 0\leqq s\leqq\delta=\log(1+\rho/|\lambda_j-\lambda_k|)$$

(with $\arg(\lambda_j-\lambda_k)=\arg(t-\lambda_j)=\eta_{jk}$), and we obtain

$$\int_{\lambda_j}^{\lambda_j+\rho e^{i\eta_{jk}}}(t-\lambda_k)^{\lambda_k'+p}y_j(t)\,dt=(\lambda_j-\lambda_k)^{\lambda_k'-p+1}\int_0^{\delta}e^{-s(p-\lambda_k'-1)}y_j(t(s))\,ds.$$

Expanding $e^{s(\lambda_k'+1)}y_j(t(s))$ in the variable s (for s close to 0, i.e., t close to λ_j) and using a standard result on the asymptotics of Laplace integrals (see, e.g., Doetsch [5]), we find that

$$\int_{\lambda_j}^{\lambda_j+\rho e^{i\eta_{jk}}}(t-\lambda_k)^{\lambda_k'+p}y_j(t)\,dt\cong(\lambda_j-\lambda_k)^{\lambda_k'-\lambda_j'-p}p^{\lambda_j'}\sum_0^{\infty}\tilde g_{kj}(m)p^{-m}$$

as $p\to\infty$ (with coefficients $\tilde g_{kj}(m)$ that can be computed from the coefficients of the expansion of $e^{s(\lambda_k'+1)}y_j(t(s))$; in particular $\tilde g_{kj}(0)=f_j(0)\Gamma(1+\lambda_j')\Gamma(-\lambda_j')$). The second integral has been shown to be *asymptotically negligible compared to the first* (by that we mean if we multiply with $(\lambda_j-\lambda_k)^{p-\lambda_k'+\lambda_j'}p^{-\lambda_j'}$, i.e., with the inverse of the explicit terms in the asymptotic obtained above, then the resulting function is asymptotically zero). Hence observing that $p^{\lambda_k'}\Gamma(p-\lambda_k')/\Gamma(p)$ has an asymptotic power series expansion in p^{-1} with leading term one, we obtain

$$\varphi_{kj}(R,p)\cong\Gamma(p)(\lambda_k-\lambda_j)^{\lambda_k'-\lambda_j'-p}p^{\lambda_j'-\lambda_k'}\sum_0^{\infty}g_{kj}(m)p^{-m}$$

with $\arg(\lambda_k-\lambda_j)=\arg(\lambda_j-\lambda_k)-\pi=\eta_{jk}-\pi$, and

$$\begin{aligned}g_{kj}(0)&=\tilde g_{kj}(0)(1-e^{2\pi i\lambda_j'})\,e^{-\pi i\lambda_j'}/(2\pi i)^2\\&=f_j(0)/(2\pi i).\end{aligned}$$

We formalize the results of the preceding discussion as follows:

PROPOSITION 3. *Let a differential equation* (0.1) *satisfying our basic assumptions and an arbitrarily selected formal fundamental solution* $H(z)$ *of the form* (0.2) *be given. Then for every fixed* k, $1 \leqq k \leqq n$, *the coefficients* $f_k(p)$ *can be represented as a linear combination of* $n-1$ *terms, each having a known asymptotic as* $p \to \infty$, *plus an error term which is asymptotically negligible compared to the other terms.*

More precisely, if $\lambda_j' \not\equiv 0 \pmod 1$ *and* $\operatorname{Re} \lambda_j' < 0$, *and if* η *is an arbitrarily fixed admissible direction (with respect to* λ_k*), then for* $R > 0$ *sufficiently large*

$$f_k(p) = \sum_{j \neq k} v_{jk} \varphi_{kj}(R, p) + I_k(R, p), \qquad p \geqq 0, \tag{2.9}$$

where $\varphi_{kj}(R, p)$ *and* $I_k(R, p)$ *are given by* (2.6) *and* (2.7), *and* v_{jk} *is the element in the* (j, k) *position of* V_ν, *with* ν *determined by*

$$\eta_\nu = \eta_{jk} = \arg(\lambda_j - \lambda_k) \in (\eta - 2\pi, \eta).$$

As $p \to \infty$, *we have (for sufficiently small* $\varepsilon > 0$)

$$I_k(R, p) = \Gamma(p) O((R - \varepsilon)^{-p}),$$

$$\varphi_{kj}(R, p) \cong \Gamma(p)(\lambda_k - \lambda_j)^{\lambda_k' - \lambda_j' - p} p^{\lambda_j' - \lambda_k'} \sum_{m=0}^{\infty} g_{kj}(m) p^{-m}, \tag{2.10}$$

with $\arg(\lambda_k - \lambda_j) = \eta_{jk} - \pi$ *and* $g_{kj}(0) = f_j(0)/(2\pi i)$.

Remark 2.1. While for the proof of Proposition 3 we required that $\lambda_j' \not\equiv 0 \pmod 1$ and $\operatorname{Re} \lambda_j' < 0$ for all $j = 1, 2, \cdots, n$, we observe that since a scalar shift of the differential equation does not change the $f_k(p)$ or the asymptotic (2.10), then these assumptions are not required for the existence of such an asymptotic for the formal coefficients $f_k(p)$ as $p \to \infty$. On the other hand, the assumption that $\lambda_j' \not\equiv 0 \pmod 1$ is necessary for the representation of $\varphi_{kj}(R, p)$ since if $\lambda_j' \equiv 0 \pmod 1$ then $y_j(t)$ does not exist.

Remark 2.2. The parameter R in (2.9) can be any sufficiently large real number. In the special case of a two-term differential equation (0.4), each function $y_j(t)$ grows no faster than a fixed power of t as $t \to \infty$ (when $|\arg t|$ is restricted to a bounded interval), hence for all p sufficiently large we may let $R \to +\infty$ in (2.6). Since $I_k(R, p) \to 0$ as $R \to +\infty$ (also for p large enough), one obtains an analogous formula to (2.9) without error term with the functions

$$\varphi_{kj}(p) = \varphi_{kj}(\infty, p)$$

having the same asymptotic (2.10). For a general differential equation (0.1), this argument does not apply since the functions $y_j(t)$ may have exponential growth as $t \to \infty$ and the loop integrals to ∞ may fail to converge. While there are several artificial ways to incorporate the error term by redefining the functions φ_{kj}, there seems to be no natural or simple way to do it. In § 5 we will obtain an analogous formula to (2.9) without error term but using some other functions, which in the case of a "two-term" equation (0.4) reduce to $\varphi_{kj}(\infty, p)$.

Remark 2.3. For each fixed k, $1 \leqq k \leqq n$, one may use the asymptotic development of the terms in (2.9) to calculate certain v_{jk} (to within any prescribed degree of accuracy). The j correspond to the highest level exponential terms, namely those $j \neq k$ for which $|\lambda_j - \lambda_k|$ is minimal. For each fixed k there is generally only one λ_j closest to λ_k, but should there be several at the same minimal distance, all the corresponding v_{jk} can be calculated in the following manner: First observe that without loss of generality we may assume A_0 is diagonal and $F_0 = I$. If $\operatorname{Re} \lambda_j'$ are all distinct, then the v_{jk} can be calculated successively starting with the j corresponding to $\max \operatorname{Re} \lambda_j'$. In doing this it is important to observe that the vectors $g_{kj}(m)$ can be explicitly calculated for each

$m \geqq 0$, hence the functions $\varphi_{kj}(R, p)(\lambda_k - \lambda_j)^{p+\lambda_j'-\lambda_k'}/\Gamma(p)$ can be calculated up to $O(p^{-N})$ for any N and p sufficiently large. In the antithetical case where $\operatorname{Re} \lambda_j'$ are all equal, then one v_{jk} can be calculated from each component by noticing that (aside from the leading terms which all have the same modulus) one term (corresponding to the unit vectors in F_0) grows like $p^{\operatorname{Re}(\lambda_j'-\lambda_k')}$ times a constant (which we want to determine) and all other terms grow at most like $p^{\operatorname{Re}(\lambda_j'-\lambda_k')-1}$. In the general case, one may put both arguments together, first calculating the v_{jk} for which $\operatorname{Re}(\lambda_j'-\lambda_k')$ is largest, and then proceeding to calculate the others as indicated.

3. Some interpolating functions for the coefficients of the formal series and their analytic and asymptotic properties. In the remaining sections we will consider a fixed differential equation satisfying only our basic assumptions and we assume that a fixed formal fundamental solution $H(z)$ of the form (0.2) has been constructed. For the purpose of extending and improving upon the representation (2.9) for the formal coefficients, we will consider now some generalizations of the associated functions $y_k(t)$, which are also natural interpolants of the formal coefficients. For $|t-\lambda_j|$ sufficiently small and all complex numbers s, we define

$$\xi_j(s, t) = \sum_{l=0}^{\infty} \frac{f_j(l)(\lambda_j - t)^{l+s-\lambda_j'-1}}{\Gamma(l+s-\lambda_j')}. \tag{3.1}$$

Observe that $\xi_j(\lambda_j'-p+1, \lambda_j) = f_j(p)$, while for each fixed $s \not\equiv \lambda_j' \pmod 1$, the function

$$2\pi i(1-e^{2\pi i(s-\lambda_j')})^{-1}\xi_j(s, t)$$

is the associated function $y_j(t)$ corresponding to the "shifted" differential equation

$$x' = (A(z) - sz^{-1}I)x$$

(if we define the powers of $(\lambda_j - t)$ and $(t-\lambda_j)$ using $\arg(\lambda_j - t) = \arg(t-\lambda_j) + \pi$). Hence, one sees (for such s) from the analytic properties of $y_j(t)$, that $\xi_j(s, t)$ may be continued analytically along every path avoiding the points $\lambda_1, \cdots, \lambda_n$, where the function generally has a branch-type singularity. To obtain a single-valued function it is therefore appropriate to cut the complex t-plane from each point $\lambda_1, \cdots, \lambda_n$ to ∞ and we choose to do this with a "star-shaped" system of cuts almost identical to what we did in § 2; however, here we do not require the cuts to be turned slightly if more than two of the λ_j's lie on the same straight line. Specifically, if j is any fixed but arbitrary integer, $1 \leqq j \leqq n$, and η is any fixed but arbitrary admissible value, we cut the t-plane from λ_j to ∞ along $\arg(t-\lambda_j) = \eta$ and from each λ_k, $k \neq j$, we cut along $\arg(t-\lambda_k) = \arg(\lambda_k - \lambda_j)$. For t not on the cut from λ_j we select

$$\arg(\lambda_j - t) \in (\eta - \pi, \eta + \pi) \tag{3.2}$$

in defining the powers of $\lambda_j - t$ and we denote the t-plane with this system of cuts and choice of the argument by $\mathcal{P}_{j,\eta}$.

For values of $s \equiv \lambda_j' \pmod 1$, the functions $\xi_j(s, t)$ are related to the derivatives or integrals of the function $\psi_j(t-\lambda_j)$ we constructed in [1, § 5]. To determine the analytic properties of $\xi_j(s, t)$ for this set of values, one could modify that discussion (in particular, see (5.9) in [1]) from the case of parallel cuts to the present star-shaped system. But the analytic continuation can also be obtained through an analogous formula (3.5) required here for another purpose, so we will give an independent argument in this case.

From the recursion relations for the coefficients $f_j(p)$ one can show that $\xi_j(s, t)$ at least formally satisfies the difference equation (0.6). To see that it is, in fact, a

solution one requires knowledge of the growth of $\xi_j(s, t)$ as $\operatorname{Re} s \to +\infty$ to establish the convergence of the infinite series. Such an estimate follows from an explicit asymptotic for $\xi_j(s, t)$ as $\operatorname{Re} s \to +\infty$.

We now formalize these analytic and asymptotic properties of the generalized associated function $\xi_j(s, t)$ as follows.

PROPOSITION 4. *In the situation described above, the function $\xi_j(s, t)$ defined by* (3.1) *is analytic in both variables for $t \in \mathcal{P}_{j,\eta}$ and all complex numbers s. Moreover, for every natural number N,*

$$\frac{\Gamma(N+s-\lambda_j')}{(\lambda_j-t)^{N+s-\lambda_j'-1}}\left\{\xi_j(s, t)-\sum_{l=0}^{N-1} f_j(l)\frac{(\lambda_j-t)^{l+s-\lambda_j'-1}}{\Gamma(l+s-\lambda_j')}\right\}=O(1) \tag{3.3}$$

as $\operatorname{Re} s \to +\infty$, *where the O-constant is uniform with respect to* $\operatorname{Im} s$ *and locally uniform with respect to t. Finally, the function $\xi_j(s, t)$ satisfies the difference equation*

$$s\xi_j(s+1, t)+t\xi_j(s, t)=\sum_{k=0}^{\infty} A_k\xi_j(s+k, t) \tag{3.4}$$

for $s \in \mathbb{C}$ and $t \in \mathcal{P}_{j,\eta}$, where the series on the right-hand side of (3.4) *is absolutely convergent due to* (3.3).

Proof. For arbitrary natural numbers $N > \operatorname{Re}(\lambda_j' - \tilde{s})$ and complex numbers s, $\tilde{s}$ satisfying $\operatorname{Re}(s-\tilde{s})>0$, one can show by integrating the expansion of $\xi_j(\tilde{s}, u)$ termwise and using a modified form of the standard beta-integral (cf. [1, p. 711]) that for $|t-\lambda_j|$ sufficiently small

$$\begin{aligned}\xi_j(s, t)=&\sum_{l=0}^{N-1} f_j(l)(\lambda_j-t)^{l+s-\lambda_j'-1}/\Gamma(l+s-\lambda_j')\\&-\int_{\lambda_j}^{t}\frac{(u-t)^{s-\tilde{s}-1}}{\Gamma(s-\tilde{s})}\left\{\xi_j(\tilde{s}, u)-\sum_{l=0}^{N-1} f_j(l)\frac{(\lambda_j-u)^{l+\tilde{s}-\lambda_j'-1}}{\Gamma(l+\tilde{s}-\lambda_j')}\right\}du\end{aligned} \tag{3.5}$$

(if one integrates along the straight line segment from λ_j to t and defines the power of $(u-t)$ according to $\arg(u-t)=\arg(\lambda_j-t)$). This formula, first established for $|t-\lambda_j|$ sufficiently small, now can be used to analytically continue $\xi_j(s, t)$ everywhere as long as the right-hand side is analytic. In particular, for $\tilde{s} \not\equiv \lambda_j' \pmod 1$ we have seen above that $\xi_j(\tilde{s}, u)$ is analytic for $u \in \mathcal{P}_{j,\eta}$, hence selecting any such $\tilde{s}$ so that $\operatorname{Re}(s-\tilde{s})>0$, we conclude that $\xi_j(s, t)$ is analytic for $t \in \mathcal{P}_{j,\eta}$ and any fixed s. (Note that since $\mathcal{P}_{j,\eta}$ is star-shaped we can reach any $t \in \mathcal{P}_{j,\eta}$ along straight line segments from λ_j.) Moreover, for any fixed value of $t \in \mathcal{P}_{j,\eta}$, the analyticity of $\xi_j(s, t)$ with respect to s also follows from (3.5) (taking any fixed $\tilde{s}$ with $\operatorname{Re}(s-\tilde{s})>0$).

To prove (3.3) we select $\tilde{s}=\lambda_j'$ in the integral on the right-hand side of (3.5) (hence N may be any natural number) and obtain by an $(N-1)$-fold partial integration that

$$\begin{aligned}&-\int_{\lambda_j}^{t}\frac{(u-t)^{s-\lambda_j'-1}}{\Gamma(s-\lambda_j')}\left\{\xi_j(\lambda_j', u)-\sum_{l=1}^{N-1}\frac{f_j(l)(\lambda_j-u)^{l-1}}{\Gamma(l)}\right\}du\\&\quad=(-1)^N\int_{\lambda_j}^{t}\frac{(u-t)^{s+N-\lambda_j'-2}}{\Gamma(s+N-\lambda_j'-1)}\left\{\left(\frac{d}{du}\right)^{N-1}\xi_j(\lambda_j', u)\right\}du\\&\quad=\frac{(\lambda_j-t)^{s+N-\lambda_j'-1}}{\Gamma(s+N-\lambda_j')}f_j(N)\\&\qquad+(-1)^{N+1}\int_{\lambda_j}^{t}\frac{(u-t)^{s+N-\lambda_j'-1}}{\Gamma(s+N-\lambda_j')}\left\{\left(\frac{d}{du}\right)^{N}\xi_j(\lambda_j', u)\right\}du\end{aligned}$$

for every natural number N and every complex s with $\operatorname{Re}(s-\lambda_j')>0$. Letting $\lambda_j - t = |\lambda_j - t|\, e^{i\varphi}$, it follows that the straight line from λ_j to t can be parametrized by

$$u - t = e^{i\varphi}|\lambda_j - t|(1-x) \quad \text{for } 0 \leqq x \leqq 1.$$

Making use of (3.5) (with $\tilde{s} = \lambda_j'$), we see that the left-hand side of (3.3) may be expressed as

$$f_j(N) + (-1)^N(\lambda_j - t)\int_0^1 (1-x)^{s+N-\lambda_j'-1}\left\{\left(\frac{d}{du}\right)^N \xi_j(\lambda_j', u)\right\} dx.$$

If K_t denotes the maximum of $\|(d/du)^N\, \xi_j(\lambda_j', u)\|$ for u between λ_j and t, then the left-hand side of (3.3) is less than or equal to

$$\|f_j(N)\| + |\lambda_j - t| K_t / \operatorname{Re}(s + N - \lambda_j')$$

which is clearly bounded (as $\operatorname{Re} s \to \infty$) by a number which is independent of $\operatorname{Im} s$ and locally uniform in t (for $t \in \mathcal{P}_{j,\eta}$). This proves (3.3).

To prove (3.4), first observe that the series converges absolutely, according to (3.3). For $|t - \lambda_j|$ sufficiently small one can prove (3.4) by substituting the expansions for the functions $\xi_j(s+k, t)$ into (3.4) and equating like powers of $\lambda_j - t$ (using the recursion formulas for the coefficients $f_j(l)$). For the other $t \in \mathcal{P}_{j,\eta}$, formula (3.4) holds by means of analytic continuation with respect to t.

Remark 3.1. The asymptotic result (3.3) can be expressed as a more familiar type of asymptotic power series in s^{-1} times certain explicit functions of s as follows: Using the usual interpretation for $\cong$, we rewrite (3.3) as

$$\xi_j(s, t) \cong \frac{(\lambda_k - t)^{s-\lambda_j'-1}}{\Gamma(s)} \sum_0^\infty f_j(l)(\lambda_j - t)^l \frac{\Gamma(s)}{\Gamma(l+s-\lambda_j')}, \qquad \operatorname{Re} s \to \infty.$$

(Note that this converges for $|t - \lambda_j|$ sufficiently small, but in general is only an asymptotic series.)

From the known asymptotic expression

$$\frac{\Gamma(s)}{\Gamma(s-\lambda_j'+l)} \cong s^{\lambda_j'} \sum_{\nu=l}^\infty c_{\nu,l}(\lambda_j') s^{-\nu} \quad \text{as } \operatorname{Re} s \to +\infty$$

(where the coefficients $c_{\nu,l}(\lambda_j')$ can be expressed in terms of Bernoulli numbers; see e.g., [16, p. 111]), we also have

$$\xi_j(s, t) \cong \frac{(\lambda_j - t)^{s-\lambda_j'-1}}{\Gamma(s)} s^{\lambda_j'} \sum_{\nu=0}^\infty s^{-\nu} \sum_{l=0}^\nu c_{\nu,l}(\lambda_j') f_j(l)(\lambda_k - t)^l. \tag{3.6}$$

Remark 3.2. M. Hukuhara [9] and R. Schäfke [17, p. 27] have obtained the asymptotic behavior of functions related to $\xi_j(s, t)$ for the two-term differential equation (0.4). Schäfke's asymptotic formula is expressed in a manner that could be considered an "asymptotic factorial series."

4. Solutions of the difference equation having a known asymptotic as $\operatorname{Re} s \to -\infty$. We have seen in the previous section that the associated functions $\xi_j(s, t)$ are solutions of (0.6) having a known asymptotic behavior as $\operatorname{Re} s \to +\infty$. For the special case of difference equations (0.5), one may construct formal solutions and use the asymptotic theory to conclude that there also exist solutions having the formal solutions as their asymptotic as $s \to \infty$ in appropriate half-planes. For the general case of (0.6), where no such formal solutions and asymptotic theory are known, it is therefore somewhat surprising that we nevertheless can construct some explicit solutions with a known

asymptotic as $\operatorname{Re} s \to -\infty$. These functions turn out to be related to the functions $y^*(t)$ we constructed in [1], § 1 and in this section we want to investigate their analytic and asymptotic properties. To construct these functions, let τ denote any fixed real number, let R denote any real number larger than a, and let $\gamma(\tau)$ be the loop-contour coming from ∞ along the ray $\arg z = \tau - 2\pi$ to the point $z_0 = R \exp[i(\tau - 2\pi)]$, then along the circle $|z| = R$ (in positive direction) to the point $z_0 e^{2\pi i}$, and back to ∞ along $\arg z = \tau$. For each fixed but arbitrary integer ν and each τ satisfying

$$\tau_\nu - \pi < \tau < \tau_{\nu+1},$$

we define

$$\xi^*(s, t; \nu) = \frac{1}{2\pi i} \int_{\gamma(\tau)} X_\nu(z) z^{-s} e^{-zt} (I - e^{2\pi i(sI - M_\nu)})^{-1} \, dz, \tag{4.1}$$

whenever the inverse matrix exists and the integral converges. Recall (§ 1) that the eigenvalues of M_ν are independent of ν and coincide with the eigenvalues of M. We label them as $\mu_1, \cdots, \mu_n$ (in some fixed, but arbitrary, ordering and repeated according to their multiplicity) and remark that $I - \exp[2\pi i(sI - M_\nu)]$ is invertible iff

$$s \not\equiv \mu_k \pmod 1, \qquad 1 \leqq k \leqq n. \tag{4.2}$$

In order to see for which values of t the integral in (4.1) converges, it is best to interpret the improper integral in the sense of its Cauchy principle value, i.e., we think of the two ray paths in $\gamma(\tau)$ as truncated at a common distance ρ and then let ρ tend to ∞. Observing that

$$X_\nu(z e^{-2\pi i})(z e^{-2\pi i})^{-s} = X_\nu(z) z^{-s} \exp[2\pi i(sI - M_\nu)],$$

it follows that (4.1) may be rewritten (using a change of variable) into a completely equivalent form as

$$\begin{aligned} \xi^*(s, t; \nu) = {} & \frac{1}{2\pi i} \int_{\gamma_1} X_\nu(z) z^{-s} e^{-zt} \, dz (I - \exp[2\pi i(sI - M_\nu)])^{-1} \\ & + \frac{1}{2\pi i} \int_{\gamma_2} X_\nu(z) z^{-s} e^{-zt} \, dz, \end{aligned} \tag{4.3}$$

where $\gamma_1 = \gamma_1(R, \tau)$ denotes the circular path $z = \operatorname{Re}^{i\varphi}$, $\tau - 2\pi \leqq \varphi \leqq \tau$, and $\gamma_2 = \gamma_2(R, \tau)$ is the ray $z = \rho e^{i\tau}$, $R \leqq \rho < \infty$. Due to the asymptotic expansion of $X_\nu(z)$ (as $z \to \infty$ in $S_\nu = S(\tau_\nu - \pi, \tau_{\nu+1})$), we see that for each j, $1 \leqq j \leqq n$, the jth column of the second integral in (4.3) converges for t satisfying

$$-\pi/2 < \arg(t - \lambda_j) + \tau < \pi/2, \tag{4.4}$$

while the first integral is an entire function of t. Therefore, $\xi^*(s, t; \nu)$ exists for every s satisfying (4.2) and t with

$$-\pi/2 < \arg(t - \lambda_l) + \tau < \pi/2, \qquad 1 \leqq l \leqq n. \tag{4.5}$$

Note that the value of the integral in (4.1) does not depend upon the choice of τ (as long as $\tau \in (\tau_\nu - \pi, \tau_{\nu+1})$ and (4.5) holds).

Letting $\xi_j^*(s, t; \nu)$ denote the jth column of $\xi^*(s, t; \nu)$ $(1 \leqq j \leqq n)$ and recalling that the Stokes' rays $\arg z = \tau_\nu$ and the critical values η_ν are related by

$$\eta_\nu + \tau_\nu = 3\pi/2 \quad \text{for every } \nu, \tag{4.6}$$

we now describe the properties of these functions in the following.

PROPOSITION 5. *Consider a differential equation* (0.1) *that satisfies our basic assumptions, a formal fundamental solution* $H(z)$, *an integer* ν, *and an index* $j(1 \leqq j \leqq n)$ *to be given, and let* $\xi_j^*(s, t; \nu)$ *denote the* j*th column of* (4.1). *Then for every* s *satisfying* (4.2), *the function* $\xi_j^*(s, t; \nu)$ *is analytic for* t *in the sector* (*on the Riemann surface of* $\log(t-\lambda_j)$)

$$\eta_{\nu+1} - 2\pi < \arg(t-\lambda_j) < \eta_\nu \tag{4.7}$$

and satisfies there the difference equation

$$s\xi_j^*(s+1, t; \nu) + t\xi_j^*(s, t; \nu) = \sum_{k=0}^{\infty} A_k \xi_j^*(s+k, t; \nu). \tag{4.8}$$

Moreover, for every (*fixed*) t *satisfying* (4.7), $\xi_j^*(s, t; \nu)$ *is meromorphic in* s *and may have poles only at the points*

$$s \equiv \mu_k \pmod 1, \qquad 1 \leqq k \leqq n. \tag{4.9}$$

Finally, if M_δ (*for sufficiently small* $\delta > 0$) *denotes the* s*-plane with* δ*-neighborhoods of the points in* (4.9) *deleted, then for every* t *satisfying*

$$\eta_{\nu+1} - 3\pi/2 < \arg(t-\lambda_j) < \eta_\nu - \pi/2$$

and every natural number N *we have*

$$\begin{aligned} &2\pi i \xi_j^*(s, t; \nu) - \sum_{l=0}^{N-1} f_j(l)\Gamma(1+\lambda_j' - s - l)(t-\lambda_j)^{l+s-\lambda_j'-1} \\ &\quad = O((t-\lambda_j)^{N+s-\lambda_j'-1}\Gamma(1+\lambda_j' - s - N)) \quad \text{as } \operatorname{Re} s \to -\infty, \quad s \in M_\delta, \end{aligned} \tag{4.10}$$

where the O-constant is locally uniform with respect to Im s *and* t.

Proof. For every $\tau \in (\tau_\nu - \pi, \tau_{\nu+1})$, we see from (4.3) that $\xi_j^*(s, t; \nu)$ (for every fixed s with (4.2)) is analytic in t in the sector given by (4.4), and by varying τ we can analytically continue $\xi_j^*(s, t; \nu)$ with respect to s. Since τ was taken arbitrarily from the interval $(\tau_\nu - \pi, \tau_{\nu+1})$, we see that $\xi_j^*(s, t; \nu)$ is analytic for t satisfying (4.7).

If t is now fixed and satisfies (4.7), while τ is selected in $(\tau_\nu - \pi, \tau_{\nu+1})$ such that (4.4) holds, then the jth column of the second integral in (4.3) is easily seen to be an entire function of s. Hence it follows from (4.3) that since

$$\int_{\gamma_1} X_\nu(z) z^{-s} e^{-zt}\, dz$$

is an entire function of s, the only singularities of $\xi_j^*(s, t; \nu)$ (with respect to s) are among the points where

$$I - \exp[2\pi i(sI - M_\nu)]$$

is not invertible, i.e., the points satisfying (4.9). At such points $\xi_j^*(s, t; \nu)$ has at worst a pole-type singularity.

To prove (4.8), we simply insert (4.1), or the equivalent form (4.3) (with s replaced by $s+k$) into the series on the right of (4.8), interchange the order of summation and integration and then integrate by parts using (0.1). To justify the interchange of the order of integration and summation and show the absolute convergence of the series, estimate (4.3) to obtain (for fixed t)

$$\xi_j^*(s+k, t; \nu) = O(R^{-k}) \quad \text{as } k \to \infty$$

and use $R > a$.

To prove (4.10), let δ, M_δ, N and t be given as in the final statement of the proposition. Note that for the t values satisfying

$$\eta_{\nu+1} - 3\pi/2 < \arg(t-\lambda_j) < \eta_\nu - \pi/2,$$

we may take $\tau = -\arg(t-\lambda_j)$ in (4.3). Defining now

$$r^{(N)}(s,t) = 2\pi i \xi_j^*(s,t;\nu) - \sum_{l=0}^{N-1} f_j(l)\Gamma(\lambda_j'+1-l-s)(t-\lambda_j)^{l+s-\lambda_j'-1}$$

and using (4.3) with the identity

$$(t-\lambda_j)^{-\alpha}\Gamma(\alpha) = \int_0^{\infty(\tau)} z^{\alpha-1} e^{-z(t-\lambda_j)}\, dz \quad \text{for } \operatorname{Re}\alpha > 0,$$

we obtain (for $-\operatorname{Re} s$ sufficiently large)

$$r^{(N)}(s,t) = r_1(s,t) + r_2^{(N)}(s,t) - r_3^{(N)}(s,t),$$

where $r_1(s,t)$ denotes the jth column of the first term on the right-hand side of (4.3),

$$r_2^{(N)}(s,t) = \int_{\operatorname{Re}^{i\tau}}^{\infty(\tau)} \left\{ x_{\nu,j}(z) - z^{\lambda_j'} e^{\lambda_j z} \sum_0^{N-1} f_j(l) z^{-l} \right\} z^{-s} e^{-zt}\, dz,$$

and

$$r_3^{(N)}(s,t) = \int_0^{\operatorname{Re}^{i\tau}} \left\{ \sum_0^{N-1} f_j(l) z^{-l} \right\} z^{\lambda_j'-s} e^{-z(t-\lambda_j)}\, dz.$$

Making the change of variable $x = z(t-\lambda_j)$ (hence $\arg x = 0$ due to the choice of τ), and using the asymptotic of $x_{\nu,j}(z)$, the jth column of the normal solution $X_\nu(z)$, for $z \to \infty$, $z \in \mathcal{S}_\nu$, we obtain

$$(t-\lambda_j)^{\lambda_j'+1-N-s} r_2^{(N)}(s,t) = \int_{R|t-\lambda_j|}^{\infty} x^{\lambda_j'-N-s} e^{-x} b(x)\, dx,$$

where $b(x)$ is bounded for $R|t-\lambda_j| \leqq x < \infty$, say by the constant K which is independent of t. Hence

$$\left\| \frac{r_2^{(N)}(s,t)(t-\lambda_j)^{\lambda_j'+1-N-s}}{\Gamma(\lambda_j'-s-N+1)} \right\| \leqq K \left| \frac{\Gamma(\operatorname{Re}(\lambda_j'-s)-N+1)}{\Gamma(\lambda_j'-s-N+1)} \right|. \tag{4.11}$$

Letting $\lambda_j' - s - N + 1 = x + iy$ with $x, y \in \mathbb{R}$, and using

$$\frac{\Gamma(x)}{\Gamma(x+iy)} \cong x^{-iy}(1+O(1)) \quad \text{as } x \to +\infty,$$

we see that the right-hand side of (4.11) is bounded by a constant that is locally uniform with respect to $\operatorname{Im} s$ and t. For the other two components of $r^{(N)}(s,t)$, one easily estimates that as $\operatorname{Re} s \to -\infty$,

$$\|r_1(s,t)\| \leqq \tilde{K}^{\operatorname{Re} s}, \qquad \|r_3^{(N)}(s,t)\| \leqq \tilde{K}^{\operatorname{Re} s}$$

for a suitably small constant $\tilde{K}$ that may be taken both independent of $\operatorname{Im} s$ and t as long as both vary locally so that $s \in M_\delta$ and $\arg(t-\lambda_j) \in (\eta_{\nu+1} - 3\pi/2, \eta_\nu - \pi/2)$. These statements then prove (4.10).

Remark 4.1. In Proposition 5 we have observed that $\xi^*(s, t; \nu)$ has possible poles (in the variable s) at the points $s \equiv \mu_k \pmod 1$, $1 \leqq k \leqq n$. To attach a meaning to the residues of $\xi^*(s, t; \nu)$ at these points, let us assume that t satisfying (4.3) is fixed and, moreover, to simplify the situation we assume that $e^{2\pi i M_\nu}$ has all distinct eigenvalues, i.e., the eigenvalues of M_ν are incongruent modulo one and are denoted by $\mu_1, \cdots, \mu_n$. Let Ω_ν be any invertible matrix satisfying

$$\Omega_\nu^{-1} \exp [2\pi i M_\nu] \Omega_\nu = \exp [2\pi i M]$$

and let $X(z) = X_\nu(z)\Omega_\nu$. Then as we observed in § 1, $X(z)z^{-M} = L(z)$ is single-valued for $a < |z| < \infty$. From (4.1), (1.4) we conclude that

$$\xi^*(s, t; \nu)\Omega_\nu = \frac{1}{2\pi i} \int_{\gamma(\tau)} L(z) z^{M-sI} e^{-zt} (I - e^{2\pi i(sI-M)})^{-1} \, dz.$$

For each fixed $j (1 \leqq j \leqq n)$ and p(integer) we find that all but the jth column of $\xi^*(s, t; \nu)\Omega_\nu$ are analytic (in s) at $s = \mu_j + p + 1$, while the jth column has a first order pole there and its residue is equal to

$$-\frac{1}{(2\pi i)^2} \int_{\gamma(\tau)} l_j(z) z^{-p-1} e^{-zt} \, dz = \frac{-1}{2\pi i} b_j(p, t),$$

if $l_j(z)$ denotes the jth column of $L(z)$, and $b_j(p, t)$ is the pth Laurent coefficient of $l_j(z)\, e^{-zt}$. This shows (under the additional assumptions made above) that the residues of $\xi^*(s, t; \nu)$ at the points $s \equiv \mu_j \pmod 1$ $(j = 1, \cdots, n)$ can be expressed in terms of the Laurent coefficients of $L(z)\, e^{-zt}$.

5. Lateral connection problems for (0.1) and (0.6) and a representation for the coefficients in a formal solution. We have seen in §§ 3 and 4 that the functions $\xi_k(s, t)$, $1 \leqq k \leqq n$, resp., $\xi_j^*(s, t; \nu)$, $1 \leqq j \leqq n$, are particular solutions of (0.6) having known asymptotics as $\operatorname{Re} s \to +\infty$, resp., $\operatorname{Re} s \to -\infty$. In the special case of (0.5) where the solution space is n-dimensional, one can conclude from the theory of difference equations that since these systems of solutions are both fundamental (this follows from the asymptotics), then these systems of functions are linearly related using certain one-periodic functions of s as coefficients (which could, in principle, also depend upon t). Equations such as (0.6), however, generally have many more than n linearly independent solutions, so no similar conclusion can be reached in the general case on the basis of the theory of difference equations alone.

We will nevertheless show in this section that these two systems of solutions for (0.6) do satisfy certain linear connection relations (5.2) and that the coefficients in these relations involve the Stokes' multipliers for the normal solutions of (0.1). We also show that (5.2) can be interpreted as yielding both a generalization and an improvement of the representation formulas for the formal coefficients we obtained in § 2 (2.5).

The function $\xi_k(s, t)$ was shown to be analytic for t in the star-shaped region $\mathcal{P}_{k,\eta}$ while the functions $\xi_j^*(s, t; \nu)$, $1 \leqq j \leqq n$, were shown to be analytic in

$$\bigcap_{j=1}^{n} \{t: \arg (t - \lambda_j) \in (\eta_{\nu+1} - 2\pi, \eta_\nu)\}.$$

For the purpose of comparing them we first need a common region where they are both defined; second, we would like (if possible) to also be able to use the asymptotic (4.10). With these objectives in mind, we now select a particular integer $\nu = \nu_j$ for each column $\xi_j^*(s, t; \nu)$ as follows.

Let k be any fixed integer, $1 \leqq k \leqq n$, and let η denote any fixed admissible value. If $j = k$ we choose $\nu_j = \nu$ satisfying

$$\eta_{\nu+1} < \eta < \eta_\nu,$$

while for each $j \neq k$ the integer ν_j is determined by

$$\eta < \eta_{\nu_j} < \eta + 2\pi \quad \text{and} \quad (j, k) \in \rho_{\nu_j}. \tag{5.1}$$

Note that ν_j depends upon both k and η but to simplify the notation we will not usually display that dependence. From the definition of ρ_{ν_j} (see [1, p. 699]) it is clear that (5.1) uniquely determines ν_j and one also sees that

$$\eta_{\nu_j} = 3\pi/2 - \tau_{\nu_j}$$

is a possible choice for $\arg(\lambda_j - \lambda_k)$, hence the region $\mathcal{P}_{k,\eta}$ can be considered part of the sector

$$\eta_{\nu_j+1} - 2\pi < \arg(t - \lambda_j) < \eta_{\nu_j}.$$

Therefore from Proposition 5 the function $\xi_j^*(s, t; \nu_j)$ (for every s satisfying (4.2)) is an analytic function in $\mathcal{P}_{k,\eta}$, $1 \leqq j \leqq n$, $j \neq k$. By similar arguments the same is true for $j = k$. To simplify the notation, we will mainly write $\xi_j^*(s, t)$ instead of $\xi_j^*(s, t; \nu_j)$, but it is important to keep in mind how these functions depend upon $\nu_j = \nu_j(k, \eta)$. We now state the connection relations between these two systems of functions as follows.

THEOREM 1. *Let a differential equation* (0.1) *satisfying our basic assumptions, a selected formal fundamental solution* $H(z)$, *any fixed index* k, $1 \leqq k \leqq n$, *and any fixed, admissible value* η *be given. Let* $\xi_k(s, t)$ *denote the analytic continuation into the cut plane* $\mathcal{P}_{k,\eta}$ *of the function* $\xi_k(s, t)$ *defined locally by* (3.1) *and let* $\xi_j^*(s, t) = \xi_j^*(s, t; \nu_j)$ *be defined by* (4.1) *with* ν_j *determined as above,* $1 \leqq j \leqq n$. *Then for every* s *satisfying* (4.2) *and* $t \in \mathcal{P}_{k,\eta}$ *we have*

$$\xi_k(s, t) = \sum_{\substack{j=1 \\ j \neq k}}^{n} v_{jk} \xi_j^*(s, t) + (1 - e^{2\pi i(s - \lambda_k')}) \xi_k^*(s, t), \tag{5.2}$$

where $v_{jk} = v_{jk}(\eta)$ *is the element in the* (j, k) *position of the Stokes' multiplier* V_{ν_j}.

Remark 5.1. This result extends Theorem 3.2.3 of R. Schäfke [17] from the "two-term" equation (0.4) (and the assumption that the position sets ρ_ν contain just a single pair) to the "general case" (0.1). Also in the case of (0.4), Schäfke has shown that essentially the same functions as $\xi_j^*(s, t)$ can be expressed as certain Mellin-type integral transforms of $\xi_j(s, t)$ (or equivalently of the $y_j(t)$) with either ray or loop paths (see [17, Def. 1.2.3 and Thm. 1.2.6]). As we remarked in § 2, in this case one can also take $R = \infty$ and $I_k(R, p) \equiv 0$ in (2.5). Then for the special choice in (5.2) of $s = \lambda_k' - p + 1$ and $t = \lambda_k$, one sees that (5.2) reduces to (2.5).

Remark 5.2. For the general case of (0.1), the function $\xi_j(s, t)$ may have exponential growth as $|t| \to \infty$, and we have no representation for $\xi_j^*(s, t)$ as an integral transform of $\xi_j(s, t)$.

Remark 5.3. Using (5.2) and the properties of $\xi_j^*(s, t)$ from Proposition 5, one sees that $\xi_j^*(s, t)$ satisfies

$$\xi_k^*(s, t) = (1 - e^{2\pi i(s - \lambda_k')})^{-1} \xi_k(s, t) + \operatorname{reg}(t - \lambda_k)$$

and

$$\xi_j^*(s, t) = \operatorname{reg}(t - \lambda_k) \quad \text{for all } k \neq j.$$

Again in the case of (0.4), since the functions $\xi_j(s, t)$ and $\xi_j^*(s, t)$, $1 \leqq j \leqq n$, are solutions of the same "shifted" associated differential equation

$$\frac{d\xi}{dt} = (A_0 - tI)^{-1}(A_1 - (s-1)I)\xi,$$

one can show that (under the additional assumption that $\lambda_j' \not\equiv 0 \pmod 1$) these conditions uniquely characterize the solutions ξ_j^*. Hence it follows (using that $2\pi i(1 - e^{2\pi i(s-\lambda_j')})^{-1}\xi_j(s, t)$ is the associated function $y_j(t)$ corresponding to the shifted equation) that

$$\xi_j^*(s, t) = (2\pi i)^{-1} y_j^*(t),$$

where $y_j^*(t)$ corresponds to the shifted differential equation (under an appropriate condition on s to guarantee its existence). It can be shown using arguments as in the proof of Proposition 6 in the next section that this is also true for (0.1).

Proof. We assume throughout that s is fixed, but arbitrary, and satisfies (4.2). First observe that it is sufficient to prove (5.2) for all t satisfying $\arg(t - \lambda_k) = \eta - \varepsilon/2$, $|t - \lambda_k| \geqq \rho$ (for some fixed and sufficiently small $\varepsilon > 0$ and some fixed and sufficiently large ρ), since then it holds for every $t \in \mathcal{P}_{k,\eta}$ by analytic continuation. If ρ is taken sufficiently large (while $\varepsilon > 0$ is fixed), one sees that for such t we have

$$\eta - \varepsilon < \arg(t - \lambda_j) < \eta, \qquad 1 \leqq j \leqq n, \tag{5.3}$$

and we will, from now on, also consider t fixed (as above).

Let ν_0 be such that $\eta_{\nu_0+m} < \cdots < \eta_{\nu_0+1}$ are precisely all the critical values in $(\eta, \eta + 2\pi)$, i.e.,

$$\tau_{\nu_0} \leqq -\eta - \pi/2 < \tau_{\nu_0+1} < \cdots < \tau_{\nu_0+m} < -\eta + 3\pi/2 \leqq \tau_{\nu_0+m+1}.$$

Then we claim that for every ν, $\nu_0 + 1 \leqq \nu \leqq \nu_0 + m$, there exists a $\tau(\nu) \in (\tau_\nu - \pi, \tau_{\nu+1})$ such that (4.5) holds with $\tau(\nu)$ in place of τ, hence $\xi^*(s, t; \nu)$ may be represented by (4.1) with $\tau = \tau(\nu)$. To see this, note that for t as above,

$$\tau + \arg(t - \lambda_j) \in (\eta + \tau - \varepsilon, \eta + \tau),$$

and we just need to show that there exists a $\tau(\nu) \in (\tau_\nu - \pi, \tau_{\nu+1})$ such that

$$[\tau(\nu) + \eta - \varepsilon, \tau(\nu) + \eta] \subset (-\pi/2, \pi/2).$$

But this follows immediately from the above inequalities.

For every $j \neq k$, $1 \leqq j \leqq n$, the corresponding ν_j lies between $\nu_0 + 1$ and $\nu_0 + m$, and the sum of those $v_{jk}\xi_j^*(s, t; \nu_j)$ for which ν_j is some fixed value ν, $\nu_0 + 1 \leqq \nu \leqq \nu_0 + m$, equals the kth column of $\xi^*(s, t; \nu)(V_\nu - I)$. Consequently, the sum of all the terms $v_{jk}\xi_j^*(s, t; \nu_j)$ $(j \neq k)$ is the kth column of

$$\sum_{\nu=\nu_0+1}^{\nu_0+m} \xi^*(s, t; \nu)(V_\nu - I)$$

$$= \sum_{\nu=\nu_0+1}^{\nu_0+m} \frac{1}{2\pi i} \int_{\gamma(\tau(\nu))} X_\nu(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_\nu)})^{-1}(V_\nu - I).$$

It easily follows from (1.2) that for every ν

$$(I - e^{2\pi i(sI - M_\nu)})^{-1} V_\nu = V_\nu (I - e^{2\pi i(sI - M_{\nu-1})})^{-1},$$

hence by means of $X_\nu(z)V_\nu = X_{\nu-1}(z)$ we obtain

$$\sum_{\nu=\nu_0+1}^{\nu_0+m} \xi^*(s, t; \nu)(V_\nu - I)$$

$$= \sum_{\nu=\nu_0}^{\nu_0+m-1} \frac{1}{2\pi i} \int_{\gamma(\tau(\nu+1))} X_\nu(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_\nu)})^{-1}$$

$$- \sum_{\nu=\nu_0+1}^{\nu_0+m} \frac{1}{2\pi i} \int_{\gamma(\tau(\nu))} X_\nu(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_\nu)})^{-1}.$$

Turning the path of integration continuously in any one of the above integrals does not change its value provided that it still converges. In view of our choice of t, all the above integrals along a $\gamma(\tau)$ converge for $\eta - \varepsilon + \tau \in (-\pi/2, \pi/2)$. Hence we find for $\nu = \nu_0+1, \cdots, \nu_0+m-1$ that corresponding terms of the above two sums are equal, and we therefore obtain

$$\sum_{\nu=\nu_0+1}^{\nu_0+m} \xi^*(s, t; \nu)(V_\nu - I)$$

$$= \frac{1}{2\pi i} \int_{\gamma(\tau(\nu_0+1))} X_{\nu_0}(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_{\nu_0})})^{-1}$$

$$- \frac{1}{2\pi i} \int_{\gamma(\tau(\nu_0+m))} X_{\nu_0+m}(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_{\nu_0+m})})^{-1}.$$

Since the largest integer ν for which $\eta_\nu > \eta$ has been denoted by ν_k, we find $\nu_k = \nu_0 + m$; hence, $\xi_k^*(s, t)$ is the kth column of

$$\xi^*(s, t; \nu_k) = \frac{1}{2\pi i} \int_{\gamma(\tau(\nu_0+m))} X_{\nu_0+m}(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_{\nu_0+m})})^{-1}.$$

Using the identities

$$(I - e^{2\pi i(sI - M_{\nu_0+m})})^{-1} e^{2\pi i(sI - \Lambda')} = e^{2\pi i(sI - \Lambda')} (I - e^{2\pi i(sI - M_{\nu_0})})^{-1}$$

and

$$X_{\nu_0+m}(z) = X_{\nu_0}(z e^{-2\pi i}) e^{2\pi i \Lambda'},$$

we find

$$\xi^*(s, t; \nu_k) e^{2\pi i(sI - \Lambda')}$$

$$= \frac{1}{2\pi i} \int_{\gamma(\tau(\nu_0+m)-2\pi)} X_{\nu_0}(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_{\nu_0})})^{-1}.$$

The proof will be completed once we show that $\xi_k(s, t)$ is the kth column of

$$\frac{1}{2\pi i} \left\{ \int_{\gamma(\tau)} - \int_{\gamma(\tilde{\tau})} \right\} X_{\nu_0}(z) z^{-s} e^{-zt}\, dz (I - e^{2\pi i(sI - M_{\nu_0})})^{-1}, \tag{5.4}$$

where $\tau = \tau(\nu_0+1)$ and $\tilde{\tau} = \tau(\nu_0+m) - 2\pi$. To see this, recall that by definition $\tau(\nu)$ was taken from

$$(\tau_\nu - \pi, \tau_{\nu+1}) \cap (-\pi/2 - \eta + \varepsilon/2, \pi/2 - \eta + \varepsilon/2) \quad \text{for } \nu = \nu_0+1, \cdots, \nu_0+m,$$

and since Stokes' rays occur in opposite pairs, then $\tau_{\nu_0} + \eta > -3/2\pi$ and $\tau_{\nu_0+1} + \eta < \pi/2$. Hence, we may take $\tau = \tau_{\nu_0+1} - \delta$ and $\tilde{\tau} = \tau_{\nu_0} - \pi + \delta$ for sufficiently small $\delta > 0$. Since

the integrals in (5.4) do not change their values when we continuously deform the paths of integration (provided that the integrals still converge), we find that we may replace $\gamma(\tau)$ by $\beta_{\nu_0}(\delta)$ and $\gamma(\tilde{\tau})$ by $\beta_{\nu_0-m}(\delta)$. (See [2, p. 156] for a definition of these paths.) Note that now the value of the integrand in the second integral is actually taken on the next higher sheet of the Riemann surface, so the change of variable $z = \tilde{z}\, e^{-2\pi i}$ in the second integral gives

$$\frac{1}{2\pi i}\int_{\beta_{\nu_0}(\delta)} X_{\nu_0}(z) z^{-s}\, e^{-zt}\, dz\, e^{2\pi i(sI - M_{\nu_0})}.$$

Hence we see that (5.4) equals

$$\frac{1}{2\pi i}\int_{\beta_{\nu_0}(\delta)} X_{\nu_0}(z) z^{-s}\, e^{-zt}\, dz. \tag{5.5}$$

In order to see that the kth column of (5.5) is $\xi_k(s, t)$, one may either consider inverting the Laplace transform representation for the normal solutions (e.g. (2.3)) and deform the path of integration, or else one may verify that the asymptotic for $X_{\nu_0}(z)$ can be substituted and term-wise integrated for $|t - \lambda_k|$ sufficiently small to obtain the expansion (3.1). Compare [2, pp. 157–159] where the latter procedure is carried out in a more general setting.

Remark 5.4. We wish to point out again that in the course of the proof we have obtained the following integral representation: For every fixed integer k, $1 \leqq k \leqq n$, every fixed integer ν, and all sufficiently small $\delta > 0$

$$\xi_k(s, t) = \frac{1}{2\pi i}\int_{\beta_\nu(\delta)} x_{\nu,k}(z) z^{-s}\, e^{-zt}\, dz$$

for t satisfying $\eta_{\nu+1} + \delta < \arg(t - \lambda_k) < \eta_\nu - \delta$ and all complex numbers s. If we restrict to t (in the above sector) with $|t|$ sufficiently large, then

$$\xi_k(s, t) = \frac{1}{2\pi i}\int_{\gamma(\tau)} x_{\nu,k}(z) z^{-s}\, e^{-zt}\, dz$$

for all τ with

$$-\pi/2 < \arg(t - \lambda_j) + \tau < \pi/2.$$

Remark 5.5. In order to solve the system of equations (5.2) for the elements v_{jk}, one must analytically continue the functions ξ_k and $\xi_j^*(1 \leqq j \leqq n)$ to some parameter values s, t with s satisfying (4.2), $t \in \mathcal{P}_{k,\eta}$, and such that

$$\det[\xi_1^*, \cdots, \xi_n^*](s, t) \neq 0.$$

It is easy to see using (4.10) that for $-\operatorname{Re} s$ sufficiently large the above determinant will be nonzero from the linear independence of the vectors $f_1(0), \cdots, f_n(0)$. For the purpose of simplifying the numerical calculations involved, it is reasonable to try to find some particularly convenient parameter values where we might either know exact values for the functions involved or be able to use their asymptotic. That there are such convenient parameter values which are also at the same time compatible with all the other requirements is not immediately obvious. But from our choice of the cuts it follows that we may specialize (5.2) in the following way to obtain a simpler relation.

For fixed k and for all $j \neq k$, $1 \leqq j \leqq n$ we observe that $\xi_j^*(s, t)$ is analytic at $t = \lambda_k$ and

$$-\arg(\lambda_k - \lambda_j) \in (\tau_{\nu_j} - \pi, \tau_{\nu_j+1})$$

according to the definition of ν_j (if consistent with the selection of $\arg(t-\lambda_j)$ for $t \in \mathcal{P}_{k,\eta}$ we take $\arg(\lambda_k-\lambda_j) \in (\eta_{\nu_j}-2\pi, \eta_{\nu_j})$). Hence we may put $t=\lambda_k$ in $\xi_j^*(s,t)$ and also use the asymptotic (4.10). Note that $\xi_k(s,t)$ does not exist when $t=\lambda_k$ for all values of s, but for the particular choice of $s=\lambda_k'-p+1$, where p is an arbitrary nonnegative integer, it follows directly from (3.1) that

$$\xi_k(\lambda_k'-p+1, \lambda_k) = f_k(p). \tag{5.6}$$

Defining now the functions $\alpha_j(s)=\alpha_j(s;k,\eta)=\xi_j^*(s,\lambda_k)$, note that the choice $s=\lambda_k'-p+1$ satisfies (4.2) if $\lambda_k' \not\equiv \mu_j \bmod 1$ $(1 \leqq j \leqq n)$. With these selections, we obtain as a special case of (5.2) the following.

COROLLARY 1. *For arbitrary, but fixed, admissible η and k, $1 \leqq k \leqq n$, assume $\lambda_k' \not\equiv \mu_j \bmod 1$, $1 \leqq j \leqq n$. Then*

$$f_k(p) = \sum_{\substack{j=1 \\ j \neq k}}^{n} v_{j,k}\alpha_j(\lambda_k'-p+1), \qquad p \geqq 0. \tag{5.7}$$

Remark 5.6. The functions $\alpha_j(\lambda_k'-p+1)$ have the asymptotic

$$\alpha_j(\lambda_k'-p+1) \cong \Gamma(p)(\lambda_k-\lambda_j)^{\lambda_k'-\lambda_j'-p}p^{\lambda_j'-\lambda_k'}\sum_{m=0}^{\infty} g_{kj}(m)p^{-m} \tag{5.8}$$

as $p \to +\infty$, where $g_{kj}(0)=f_j(0)/(2\pi i)$ and in defining the nonintegral power of $(\lambda_k-\lambda_j)$ we take $\arg(\lambda_k-\lambda_j)$ as above. If the asymptotic (5.8) for each of the $\alpha_j(\lambda_k'-p+1)$ is substituted into (5.7), we see that the right-hand side has a multi-level asymptotic representation as $p \to +\infty$ where the different levels are distinguished by the different exponential orders of growth $(\lambda_k-\lambda_j)^{-p}$ for $j \neq k$. For each fixed value of k, generally one level dominates and the corresponding coefficient $v_{j,k}$ can be calculated directly as a limit of the formal coefficients $f_k(p)$ after dividing out the main terms in the asymptotic. This extends results of Jurkat, Lutz and Peyerimhoff [11], who treated the case $n=2$, and R. Schäfke [17] who treated the "two-term" equation (0.4) for general n. Also see [7], where an equivalent result is obtained for a scalar second order equation that is already a special case of the results in [11].

To calculate the remaining coefficients $v_{j,k}$ $(n>2)$ one could either try to use (5.7) by calculating appropriate values for the functions $\alpha_j(\lambda_k'-p+1)$ or one could return to (5.2) and make another selection for t that would make other terms in the asymptotic dominate. For certain configurations of the eigenvalues $\lambda_1, \cdots, \lambda_n$ it is possible to choose appropriate t-values for which the asymptotic (4.10) is still valid and preselected terms will dominate the asymptotic. In using (5.2) it is of course necessary to perform the analytic continuation of the functions $\xi_k(s,t)$ possibly outside of the circle of convergence of its local power series expansion. If this is done for certain values of s, then either the difference equation (0.6) or the integral representation (3.5) may be used to calculate values of the analytic continuation of $\xi_j(s,t)$ at a set of points with $\operatorname{Re} s \to -\infty$.

In any case, one may view the functions $\xi_k(s,t)$ as a natural interpolation of the coefficients $f_k(p)$ of the formal solution vectors, and by performing the analytic continuation of these functions to certain values of the parameters one may be able to solve (5.2) for the lateral connection coefficients.

6. The central connection problem and representation of solutions in Floquet form. Our goals in this section and the next one are to investigate the relations between the Laurent coefficients of the single-valued part of a solution (vector or matrix) expressed in Floquet form (0.3) and the central connection coefficients linking that

solution with the normal solutions. The main formula expresses the Laurent coefficients (of a shifted equation) as a linear combination of the associated functions (and possibly their derivatives) evaluated at certain points times certain central connection coefficients. Two types of applications are indicated: if the Laurent coefficients would be known, the formulas may be used to calculate central connection coefficients; if the Stokes' multipliers would be known, then the Laurent coefficients of a solution in Floquet form can be calculated.

For our purposes it is convenient to introduce some different systems of cuts from the ones we used previously for the lateral connection problem. We wish to point out to the reader that for *any* fixed system of cuts from the points $\lambda_1, \cdots, \lambda_n$ and any fixed selection of $\arg(t-\lambda_j)$ for t close to λ_j, associated functions may be defined locally by (3.1) and continued analytically into all of the simply connected domain determined by the cuts. The values of the resulting continuations at various points may depend, however, on the location of the cuts. The main reason for preferring certain systems of cuts over others involves the ability to easily identify quantities arising from the continuations of the associated functions with some natural quantities corresponding to the differential equation, such as Stokes' multipliers.

Let λ denote any fixed, but arbitrary, complex number satisfying $\lambda \neq \lambda_j$, $1 \leqq j \leqq n$, and let θ denote any fixed, but arbitrary, real number satisfying

$$\theta \not\equiv \arg(\lambda_j - \lambda) \pmod{2\pi}, \qquad 1 \leqq j \leqq n.$$

For each such pair (λ, θ), which we henceforth call *admissible*, we cut the t-plane from λ to ∞ along $\arg(t-\lambda) = \theta$, and for t not on the cut we define

$$\arg(t-\lambda) \in (\theta - 2\pi, \theta).$$

Denoting $\theta_j = \arg(\lambda_j - \lambda) \in (\theta - 2\pi, \theta)$, $1 \leqq j \leqq n$, we call a λ-value *generic* if all the θ_j are *distinct*. In such a case we also make a cut from each point λ_j to λ along the straight line segment and denote the t-plane with these cuts by $\mathcal{P}_{\lambda,\theta}$. (Later on it will be important to discuss nongeneric values of λ, for which the cuts are then somewhat more complicated to arrange.) In order to define the associated functions $\xi_j(s, t)$ for $t \in \mathcal{P}_{\lambda,\theta}$, it is sufficient to select a branch of $\arg(\lambda_j - t)$, and we do this here by requiring

$$\theta_j < \arg(\lambda_j - t) < \theta_j + 2\pi, \qquad 1 \leqq j \leqq n.$$

If for a certain (fixed) generic value λ, and a given θ there exists a (necessarily unique) integer ν such that

$$\eta_{\nu+1} < \theta < \eta_\nu \quad \text{and} \quad \eta_{\nu+1} < \theta_j + \pi < \eta_\nu, \qquad 1 \leqq j \leqq n, \tag{6.1}$$

then the system of cuts is called *almost parallel*. (Here the numbers η_ν, the so-called *critical values*, are all numbers of the form $\arg(\lambda_j - \lambda_k)$, $j \neq k$; see [1, § 3.1].) This property depends both on the geometry of the numbers $\lambda_1, \cdots, \lambda_n$ as well as the choice of λ and θ. We observe that for any (fixed) set of distinct complex numbers $\lambda_1, \cdots, \lambda_n$ and any θ that is not one of the critical values, there always exists a generic λ, $|\lambda|$ sufficiently large, such that the system of cuts is almost parallel.

If $X(z)$ is any fundamental solution matrix for (0.1), we define

$$\Phi(s, t; \theta) = \Phi(s, t) = \frac{1}{2\pi i} \int_{\gamma(\tau)} X(z) z^{-s} e^{-zt}\, dz, \tag{6.2}$$

with $\gamma(\tau)$ as in (4.1), $\tau = \pi - \theta$, $s \in \mathbb{C}$, arbitrary, $t \in \mathcal{P}_{\lambda,\theta}$, and $|t|$ sufficiently large. Clearly the integral converges when $|t|$ is sufficiently large and $\arg t = \theta - \pi$ (using the asymptotic behavior of solutions). From (1.4) and the integral representation of $\xi_j(s, t)$

derived in the proof of Theorem 1, one sees that the columns of $\Phi(s, t)$ may be expressed as certain linear combinations of appropriate analytic continuations of $\xi_1(s, t), \cdots, \xi_n(s, t)$ in $\mathcal{P}_{\lambda,\theta}$. It follows that $\Phi(s, t)$ may be continued analytically (in t) to all of $\mathcal{P}_{\lambda,\theta}$. Moreover, since the vectors $\xi_1(s, t), \cdots, \xi_n(s, t)$ are all solutions of (0.6) we see that the columns of $\Phi(s, t)$ are also solutions of (0.6). We next consider the behavior of the columns $\varphi_j(s, t)$ of $\Phi(s, t)$ near the singularities $\lambda_1, \cdots, \lambda_n$ and see how the central connection factors enter into these relations. This result we now state as follows.

PROPOSITION 6. *Let a differential equation* (0.1) *satisfying our basic assumptions and a fixed formal fundamental solution be given, consider an admissible pair* (λ, θ) *for which the corresponding system of cuts is almost parallel, and let* ν *denote the unique integer satisfying* (6.1). *Then the following statements hold*:

(a) *There exists a unique matrix* $C(s) = [c_{kj}(s)]$ *of entire, one-periodic functions of* s *satisfying*

$$\xi_j(s, t) = \xi_k(s, t)(1 - e^{2\pi i(s-\lambda_k')})^{-1} c_{kj}(s) + \operatorname{reg}(t - \lambda_k) \tag{6.3}$$

for $t \in \mathcal{P}_{\lambda,\theta}$, $s \not\equiv \lambda_k' \pmod 1$, $1 \leqq j, k \leqq n$. *Moreover,* $C(s)$ *can be expressed in terms of the Stokes' multipliers as*

$$C(s) = V_\nu \cdots V_{\nu-\mu+1} - e^{2\pi i(sI-\Lambda')}(V_{\nu-\mu} \cdots V_{\nu-m+1})^{-1}, \qquad \mu = m/2. \tag{6.4}$$

(b) *If* $X(z)$ *is a fundamental solution for* (0.1) *and* $\Phi(s, t) = [\varphi_1(s, t), \cdots, \varphi_n(s, t)]$ *is defined by* (6.2), *then there exists a unique matrix* $\Delta(s) = [\delta_{kj}(s)]$ *of entire, one-periodic functions of* s *such that*

$$\varphi_j(s, t) = \xi_k(s, t)(1 - e^{2\pi i(s-\lambda_k')})^{-1} \delta_{kj}(s) + \operatorname{reg}(t - \lambda_k) \tag{6.5}$$

for $t \in \mathcal{P}_{\lambda,\theta}$, $s \not\equiv \lambda_k' \pmod 1$, $1 \leqq j, k \leqq n$. *Finally, if* $X(z) = X_\nu(z)\Omega_\nu$, *where* $X_\nu(z)$ *denotes the* ν*th normal solution, then*

$$\Delta(s) = C(s)\Omega_{\nu-\mu} = \Omega_\nu(I - e^{2\pi i(sI-M)}), \tag{6.6}$$

where $e^{2\pi iM}$ *denotes the circuit factor for* $X(z)$.

Remark 6.1. In [1] we proved a formula (see Theorem 1, § 4.2) corresponding to (6.3) for the associated functions $y_j(t)$ in the case of parallel cuts and it is readily seen that the same proof may be used to show existence (and uniqueness) of $C(s)$, $\Delta(s)$ satisfying (6.3), (6.5) for all systems of cuts that occur in §§ 6 and 7. The expressions (6.4), (6.6), however, generally depend on the particular system of cuts (since the analytic continuations of $\xi_j(s, t)$, $\varphi_j(s, t)$ depend upon the location of the cuts). We will discuss the dependence of $C(s)$, $\Delta(s)$ on the selection of the cuts in § 7, and it may be seen from the proof given there that even the existence of these matrices may be obtained along with expressions explaining how they change, provided the existence was proven in a particular situation, say the case of almost parallel cuts. To keep the statements in Lemma 1 (§ 7) as simple as possible, we will, however, take the existence of $C(s)$, $\Delta(s)$ for granted.

Proof. The uniqueness of $C(s)$, $\Delta(s)$ is immediate since for every j, $k \in \{1, 2, \cdots, n\}$, equation (6.3), resp. (6.5), holds for at most one entire function $c_{kj}(s)$, resp. $\delta_{kj}(s)$; observe that for $s \not\equiv \lambda_k' \pmod 1$ no nonzero constant multiple of $\xi_k(s, t)$ can be single-valued at λ_k.

To show (6.4), note that as a consequence of (6.1), for every j, $1 \leqq j \leqq n$, none of the cuts in $\mathcal{P}_{\lambda,\theta}$ intersects the sector

$$\eta_{\nu+\mu+1} = \eta_{\nu+1} - \pi < \arg(t - \lambda_j) < \eta_\nu - \pi = \eta_{\nu+\mu}, \tag{6.7}$$

and from the Remark 5.4, we have for t as above and arbitrary complex s,

$$\xi_j(s, t) = \frac{1}{2\pi i} \int_{\beta_{\nu-\mu}(\delta)} x_{\nu-\mu,j}(z) z^{-s} e^{-zt}\, dz.$$

For $j = 1, 2, \cdots, n$, note that the sectors (6.7) have a nonempty intersection and for such t we may replace $\beta_{\nu-\mu}(\delta)$ by $\gamma(\tau)$ with $\tau = \pi - \theta$ (θ as in (6.1)) to obtain

$$\xi(s, t) = \frac{1}{2\pi i} \int_{\gamma(\tau)} X_{\nu-\mu}(z) z^{-s} e^{-zt}\, dz. \tag{6.8}$$

Therefore from (6.2) and $X(z) = X_{\nu-\mu}(z)\Omega_{\nu-\mu}$ we obtain

$$\Phi(s, t) = \xi(s, t)\Omega_{\nu-\mu} \quad \text{for } t \in \mathcal{P}_{\lambda,\theta} \text{ and } s \in \mathbb{C}. \tag{6.9}$$

Using (6.9) we see that statement (a) in Proposition 6 implies (b) and the first equality in (6.6). From (6.4), (1.1), (1.5) and (1.6), it is easy to establish the second equality in (6.6). (Note that (1.1), (1.5), (1.6) do not depend upon M being in Jordan form.) To prove (a) and (6.4) we will show that they hold under the additional assumption (4.2) and then extend the formulas to the discrete set of values $s \equiv \mu_k \pmod 1$ using (3.5) and arguing as in the proof of Lemma 2′ in [1, p. 711].

Now assuming (4.2), let τ be as above and define

$$\xi^*(s, t) \equiv \xi^*(s, t; \nu)$$

by (4.1) (for $t \in \mathcal{P}_{\lambda,\theta}$ and satisfying (4.5)). Using (4.1), (6.8) and $X_{\nu-\mu} = X_\nu V_\nu \cdots V_{\nu-\mu+1}$, we obtain

$$\begin{aligned} \xi(s, t) &= \xi^*(s, t)(I - e^{2\pi i(sI - M_\nu)}) V_\nu \cdots V_{\nu-\mu+1} \\ &= \xi^*(s, t)\tilde{C}(s) \end{aligned} \tag{6.10}$$

with $\tilde{C}(s)$ defined by (6.4). It remains to show that these quantities are the same as in (6.3). Using Proposition 5 we conclude that

$$\xi^*(s, t) = [\xi_1^*(s, t), \cdots, \xi_n^*(s, t)]$$

can be analytically continued (with respect to t) into $\mathcal{P}$ (again use (6.1)) and

$$\xi_j^*(s, t) = \operatorname{reg}(t - \lambda_k), \qquad j \neq k, \quad 1 \leqq j, k \leqq n.$$

Hence for each such j, $1 \leqq j \leqq n$, we have by (6.10)

$$\xi_j(s, t) = \xi_k^*(s, t)\tilde{c}_{kj}(s) + \operatorname{reg}(t - \lambda_k), \qquad 1 \leqq k \leqq n,$$

which implies, in particular, for $j = k$ and $s \not\equiv \lambda_k' \pmod 1$ that

$$\xi_k^*(s, t) = \xi_k(s, t)(1 - e^{2\pi i(s - \lambda_k')})^{-1} + \operatorname{reg}(t - \lambda_k), \qquad 1 \leqq k \leqq n.$$

Here we have used that $V_\nu \cdots V_{\nu-\mu+1}$ and $(V_{\nu-\mu} \cdots V_{\nu-m+1})^{-1}$ both have ones along the diagonal, hence $c_{kk}(s) = 1 - e^{2\pi i(s - \lambda_k')}$. These formulas imply $\tilde{C} = C$, hence we obtain statement (a) with $C(s)$ as in (6.4), and this completes the proof.

Now consider a fundamental solution matrix in Floquet form, say $X(z) = L(z)z^M$, where M is an upper triangular matrix in Jordan canonical form. Since by an appropriate permutation of the columns of $X(z)$ (which corresponds to a permutation similarity of M) we can arrange any block of M to come first, and since it is notationally convenient to consider just the columns corresponding to the first block, we now restrict ourselves to that case without loss of generality. Let μ, resp. n_1, denote the eigenvalue, resp. dimension, of the first block and let $l_j(z)$, $1 \leqq j \leqq n_1$, denote the first n_1 columns of $L(z)$.

If (λ, θ) is an admissible pair with λ generic, let $t \in \mathcal{P}_{\lambda,\theta}$ with $\arg t = \theta - \pi$. For such t with $|t|$ sufficiently large (and $\gamma(\tau)$, τ as in (6.2)) we define

$$b_j(s,t) = \frac{1}{2\pi i} \int_{\gamma(\tau)} l_j(z) z^{-s-1} e^{-zt}\, dz, \qquad 1 \leqq j \leqq n_1, \tag{6.11}$$

which converges for all complex numbers s. (Note that $b_j(s,t)$ also depends on λ, θ but we do not explicitly display that here.) To discuss the analytic behavior in the t-variable, write

$$L(z) = X(z) z^{-M} = X(z)\, e^{-M \log z}$$

from which we obtain (with x_j denoting the jth column of X)

$$l_j(z) = z^{-\mu} \sum_{k=0}^{j-1} \frac{(-\log z)^k}{k!} x_{j-k}(z), \qquad 1 \leqq j \leqq n_1.$$

Inserting this into (6.11) and comparing with (6.2) we have

$$b_j(s,t) = \sum_{k=0}^{j-1} \frac{1}{k!} \left(\frac{\partial}{\partial s}\right)^k \varphi_{j-k}(\mu + s + 1, t), \qquad 1 \leqq j \leqq n_1. \tag{6.12}$$

This holds first for $|t|$ sufficiently large, $\arg t = \theta - \pi$, and all $s \in \mathbb{C}$, but using the analyticity of $\Phi(s,t)$ for all s and $t \in \mathcal{P}_{\lambda,\theta}$ we see that $b_j(s,t)$ can be continued analytically for all $t \in \mathcal{P}_{\lambda,\theta}$.

If s is an integer, it follows from (6.11) that $b_j(s,t)$ is the sth Laurent coefficient of $l_j(z)\, e^{-zt}$ (for fixed t); hence $b_j(s,t)$ is an entire function of t when s is an integer. In particular, we may select $t = \lambda$ and define

$$b_j(p) = b_j(p, \lambda)$$

for every integer p and $j = 1, 2, \cdots, n_1$. (Note that whereas $b_j(s,t)$ may depend on θ, $b_j(p)$ is clearly independent of θ.) As a consequence of interpreting $b_j(p)$ as the pth Laurent coefficient of $l_j(z)\, e^{-z\lambda}$ we obtain

$$l_j(z)\, e^{-z\lambda} = \sum_{p=-\infty}^{+\infty} b_j(p) z^p, \qquad 1 \leqq j \leqq n_1,$$

with the series converging absolutely for $|z| > a$ and uniformly for $a < R_1 \leqq |z| \leqq R_2 < +\infty$. Since $l_j(z)\, e^{-z\lambda}$ is at most of (exponential) order one and finite type as $z \to \infty$ then it is permissible to substitute the series into (6.11) and termwise integrate. To see this note that $\gamma(\tau)$ consists of two ray paths to ∞ along $\arg z = \tau$, $\tau - 2\pi$ and a circular segment. Termwise integration along the circular segment is justified by uniform convergence while termwise integration along the rays is justified for $|t - \lambda|$ sufficiently large and $\arg(t - \lambda) = \theta$ by the dominated convergence theorem. This yields

$$b_j(s,t) = \sum_{-\infty}^{+\infty} b_j(p)(\lambda - t)^{s-p} / \Gamma(1 + s - p), \qquad 1 \leqq j \leqq n_1, \tag{6.13}$$

for arbitrary s and t as above. But then the series automatically converges for all $|t - \lambda|$ sufficiently large, $t \in \mathcal{P}$ by analyticity of $b_j(s,t)$ and we have (6.13) for all such t (provided $\arg(\lambda - t) \in (\theta - \pi, \theta + \pi)$).

In order to establish our main formula which represents $b_j(p)$ in terms of the associated functions, we need an interpretation for $\xi_j(s, \lambda)$ (as well as its derivatives with respect to s) at points $s = p + \mu + 1$, where p is an integer and μ is the eigenvalue of M for the first block. To do this, we define

$$\left(\frac{\partial}{\partial s}\right)^k \xi_j(s, \lambda)$$

to be the analytic continuation of $(\partial/\partial s)^k \xi_j(s, t)$ along the segment $\arg(\lambda_j - t) = \theta_j$ ($k \geqq 0$, $1 \leqq j \leqq n_1$), which can always be performed all the way to $t = \lambda$ since we have assumed λ to be generic. We now state the representation result as follows.

THEOREM 2. *Under the same assumptions and notation as in Proposition* 6, *let* $X(z) = L(z) z^M$ *denote a fundamental solution matrix for* (0.1) *in Floquet form. Moreover, let* $b_j(p)$ *denote the pth Laurent coefficient of the vector* $l_j(z)\, e^{-z\lambda}$, *where* $l_j(z)$ *denotes the jth column of* $L(z)$. *Then* $b_j(p)$ *can be represented as a linear combination of the associated functions* (*and their derivatives with respect to* s) *with coefficients from the central connection matrix* Ω_ν. *More precisely, if* M *is in upper triangular Jordan canonical form with first block having the eigenvalue* μ *and size* n_1, *then for every integer* p *and* $j = 1, 2, \cdots, n_1$,

$$b_j(p) = \sum_{q=0}^{j-1} \frac{1}{q!} \sum_{k=1}^{n} \omega_{k,j-q}^{(\nu)} \left(\frac{\partial}{\partial s}\right)^q \xi_k(s, \lambda)\Big|_{s=p+\mu+1}, \tag{6.14}$$

where $\Omega_\nu = (\omega_{jk}^{(\nu)})$.

Proof. From (6.13) we have for each fixed, but arbitrary, integer p, for each fixed $s \not\equiv 0 \pmod 1$, and for R sufficiently large

$$b_j(p) = -\frac{\Gamma(1+s-p)}{2\pi i} \int_{|t-\lambda|=R} (\lambda - t)^{p-s-1} b_j(s, t)\, dt, \qquad 1 \leqq j \leqq n_1.$$

For those s such that $p > \operatorname{Re} s$, we may deform the contour into a sum of loops $\alpha_k(\lambda)$, where $\alpha_k(\lambda)$ is a simple closed path beginning at λ and proceeding toward λ_k close to the right border of the cut $\arg(t - \lambda) = \theta_k$, encircling λ_k, and proceeding back to λ close to the left border of the cut. No other λ_l, $l \neq k$, should lie on $\alpha_k(\lambda)$ or in the interior of the region bounded by it.

Then using also (6.12) we obtain

$$\begin{aligned} &-\frac{2\pi i}{\Gamma(1+s-p)} b_j(p) \\ &\quad = \sum_{q=0}^{j-1} \frac{1}{q!} \sum_{k=1}^{n} \int_{\alpha_k(\lambda)} (\lambda - t)^{p-s-1} \left(\frac{\partial}{\partial s}\right)^q \varphi_{j-q}(\mu + s + 1, t)\, dt. \end{aligned} \tag{6.15}$$

For fixed, but arbitrary, p, λ, q, j and k as above, and arbitrary complex numbers s and $\tilde{s}$, $s \not\equiv 0 \pmod 1$, consider the integral

$$\begin{aligned} &-\frac{\Gamma(1+s-p)}{2\pi i} \int_{\alpha_k(\lambda)} (\lambda - t)^{p-s-1} \left(\frac{\partial}{\partial \tilde{s}}\right)^q \varphi_{j-q}(\tilde{s}, t)\, dt \equiv I(s, \tilde{s}) \\ &\quad = \left(\frac{\partial}{\partial \tilde{s}}\right)^q \left[-\frac{\Gamma(1+s-p)}{2\pi i} \int_{\alpha_k(\lambda)} (\lambda - t)^{p-s-1} \varphi_{j-q}(\tilde{s}, t)\, dt \right]. \end{aligned}$$

Using (6.5) and Cauchy's theorem, one may replace $\varphi_{j-q}(\tilde{s}, t)$ by $\xi_k(\tilde{s}, t)(1 - e^{2\pi i(\tilde{s} - \lambda_k')})^{-1} \delta_{k, j-q}(\tilde{s})$ in the above integral (without changing its value). We claim that

then it follows that

$$I(s,\tilde{s}) = (1-e^{2\pi i s})^{-1}\left(\frac{\partial}{\partial\tilde{s}}\right)^q [\delta_{k,j-q}(\tilde{s})\xi_k(p-s+\tilde{s},\lambda)]. \tag{6.16}$$

To see this, first assume that in addition $\operatorname{Re}(\tilde{s}-\lambda_k')>0$. Then the integral converges at λ_k and the path $\alpha_k(\lambda)$ may be deformed to consist of two paths along the line segment from λ to λ_k, resp., from λ_k to λ, along the right, resp., left, borders of the cut. Using (3.1) one sees that the integrand on the left border is $e^{2\pi i(\tilde{s}-\lambda_k')}$ times the value on the right border, hence

$$I(s,\tilde{s}) = \left(\frac{\partial}{\partial\tilde{s}}\right)^q\left[-\frac{\Gamma(1+s-p)}{2\pi i}\delta_{k,j-q}(\tilde{s})\int_\lambda^{\lambda_k}(\lambda-u)^{p-s-1}\xi_k(\tilde{s},u)\,du\right],$$

where the integration takes place along the border with $\arg(\lambda_k-u)=\theta_k$, hence $\arg(\lambda-u)=\theta_k+\pi$. Comparing the above integral with (3.5) (taking $t=\lambda$ and $N=0$) we obtain

$$I(s,\tilde{s}) = \left(\frac{\partial}{\partial\tilde{s}}\right)^q\left[\frac{\Gamma(1+s-p)\Gamma(p-s)}{2\pi i}e^{i\pi(p-s)}\delta_{k,j-q}(\tilde{s})\xi_k(p-s+\tilde{s},\lambda)\right]$$

from which (6.16) immediately follows under the additional assumption that $\operatorname{Re}(\tilde{s}-\lambda_k')>0$. But since both $I(s,\tilde{s})$ and the right-hand side of (6.16) are entire functions of $\tilde{s}$, then (6.16) holds for all $\tilde{s}$. We also remark that in the same manner one can use (3.5) to show that

$$\begin{aligned}&-\frac{\Gamma(1-s)}{2\pi i}[1-e^{2\pi i(\tilde{s}-\lambda_k')}]^{-1}\int_{\alpha_k(\lambda)}(\lambda-t)^{s-1}\xi_k(\tilde{s},t)\,dt\\ &\quad=(1-e^{-2\pi i s})^{-1}\xi_k(s+\tilde{s},\lambda)\end{aligned} \tag{6.17}$$

for all $\tilde{s}\not\equiv\lambda_k'\pmod 1$ and $\operatorname{Re}s>0$, $s\not\equiv 0\pmod 1$.

Now returning to (6.15) and using (6.16) together with (6.6) note that in the case $j=1$ (hence $q=0$) we obtain (6.14). To complete the proof in the remaining cases we obtain from (6.15) and (6.16)

$$b_j(p) = \beta_j(p-s,\mu+s+1)(1-e^{2\pi i s})^{-1} \qquad (1\leqq j\leqq n_1)$$

with β_j defined by

$$\begin{aligned}\beta_j(p-s,\tilde{s}) &= \sum_{q=0}^{j-1}\frac{1}{q!}\sum_{k=1}^{n}\left(\frac{\partial}{\partial\tilde{s}}\right)^q[\delta_{k,j-q}(\tilde{s})\xi_k(p-s+\tilde{s},\lambda)]\\ &= \sum_{\tau=0}^{j-1}\frac{1}{\tau!}\sum_{k=1}^{n}\left(\frac{\partial}{\partial\tilde{s}}\right)^\tau\xi_k(p-s+\tilde{s},\lambda)\sum_{q=0}^{j-\tau-1}\frac{1}{q!}\left(\frac{\partial}{\partial\tilde{s}}\right)^q\delta_{k,j-\tau-q}(\tilde{s})\end{aligned}$$

(use Leibnitz's Rule to differentiate the product). Computing the first Jordan block of $e^{-2\pi i M}$ explicitly, we conclude from (6.6)

$$\delta_{kj}(\tilde{s}) = \omega_{kj}^{(\nu)} - e^{2\pi i(\tilde{s}-\mu)}\sum_{\sigma=0}^{j-1}\frac{(-1)^\sigma(2\pi i)^\sigma}{\sigma!}\omega_{k,j-\sigma}^{(\nu)} \qquad (1\leqq k\leqq n,\ 1\leqq j\leqq n_1),$$

and after some calculation (using a simple binomial identity) we obtain

$$\sum_{q=0}^{j-\tau-1}\frac{1}{q!}\left(\frac{\partial}{\partial\tilde{s}}\right)^q\delta_{k,j-\tau-q}(\tilde{s}) = (1-e^{2\pi i(\tilde{s}-\mu)})\omega_{k,j-\tau}^{(\nu)} \qquad (1\leqq k\leqq n,\ 1\leqq j-\tau\leqq n_1).$$

This implies for $j=1,\cdots,n_1$

$$\beta_j(p-s,\tilde{s}) = (1-e^{2\pi i(\tilde{s}-\mu)})\sum_{\tau=0}^{j-1}\frac{1}{\tau!}\sum_{k=1}^{n}\omega_{k,j-\tau}^{(\nu)}\left(\frac{\partial}{\partial\tilde{s}}\right)^\tau\xi_k(p-s+\tilde{s},\lambda).$$

This is easily seen to complete the proof.

Remark 6.2. For $j=1$, (6.14) reduces to

$$b_1(p)=\sum_{k=1}^{n}\omega_{k1}^{(\nu)}\xi_k(p+\mu+1,\lambda)$$

and from the asymptotic (3.3) we see that for p (positive) sufficiently large

$$\det[\xi_1(p+\mu+1,\lambda),\cdots,\xi_n(p+\mu+1,\lambda)]\neq 0. \tag{6.18}$$

Hence if the vector $b_1(p)$ would be known for one such value of p and if the analytic continuation of the $\xi_k(s,t)$ is performed, then the first column of Ω_ν can be calculated by solving the system of linear equations. For $j=2$ we obtain

$$b_2(p)-\sum_{k=1}^{n}\omega_{k1}^{(\nu)}\frac{\partial}{\partial s}\xi_k(s,\lambda)\Big|_{s=p+\mu+1}=\sum_{k=1}^{n}\omega_{k2}^{(\nu)}\xi_k(p+\mu+1,\lambda)$$

and if the left-hand side is now considered as known for some p satisfying (6.18), then the second column of Ω_ν may be calculated. Continuing in this manner (through this block of M and corresponding formulas for the others) one sees that if for each eigenvalue μ of M, (6.18) would hold for a value of p and the corresponding coefficients $b_j(p)$ would be known, and if the appropriate analytic continuations of the associated functions and their derivatives would be performed, then the central connection factor Ω_ν can be calculated.

We also remark that using the asymptotic (3.3) for $\xi_k(p+\mu+1,\lambda)$ as $p\to+\infty$ one sees that $b_1(p)$ is a sum of n terms, each having an asymptotic that consists of the common factor $1/\Gamma(p)$ times (different) exponentials in powers of p, and asymptotic power series in $1/p$. Using these asymptotics, one sees that if we would know the sequence $\{b_1(p)\}$ asymptotically in this sense as $p\to+\infty$ then (generally) one element of the column $\omega_{k1}^{(\nu)}$, $1\leqq k\leqq n$, can be calculated as a limit, corresponding to the value of k for which $|\lambda_k-\lambda|$ is the smallest.

The other columns $b_j(p)$, $j>1$, also have asymptotic representations as $p\to+\infty$ that consist of a sum of terms of the above type times various powers of $\log p$ (up to the $(j-1)$st power). This is obtained from (3.3) by differentiating with respect to s and using

$$\frac{\Gamma'(p)}{\Gamma(p)}=\Psi(p)\cong\log p-\frac{1}{2p}-\sum_{1}^{\infty}B_{2m}p^{-2m}/2m.$$

Remark 6.3. Using formula (6.10) and the knowledge of the asymptotic (4.10) for the functions $\xi_j^*(s,t)$, one could also try using (6.14) to obtain an asymptotic representation for $b_j(p)$ as $p\to-\infty$. But the leading terms of such an asymptotic would consist of $\Gamma(-p)$ times exponentials and powers of p whereas we know from the differential equation (0.1) that the Laurent coefficients can grow at most exponentially as $p\to-\infty$ since the series converges for $|z|>a$. Hence the leading terms all cancel leaving no apparent asymptotic structure for the Laurent coefficients corresponding to negative powers of z. This is indicated by the fact that, as a consequence of Birkhoff's reduction theorem, every equation (0.1) may be transformed into an equation (0.4) for which all Laurent coefficients with sufficiently small negative indices have to vanish identically.

7. The nongeneric cases; dependence of the matrices $C(s)$ and $\Delta(s)$ on the selection of λ. In the previous section we have considered the case of generic λ (in fact only the situation with almost parallel cuts) and from Remark 6.2 one may see that it may be advantageous from a computational standpoint to select several different generic

values of λ in a single problem. This makes it necessary to understand how the quantities appearing in the previous section depend upon the choice of λ, and in particular, to see what happens in the case of nongeneric values of λ.

Again consider an admissible pair (λ, θ) with the t-plane cut from λ to ∞ along $\arg(t-\lambda)=\theta$, and define $\arg(t-\lambda)\in(\theta-2\pi, \theta)$ for t not on the cut. If $\theta_j = \arg(\lambda_j-\lambda)\in(\theta-2\pi, \theta)$, let σ denote any permutation of $1, 2, \cdots, n$ such that

$$\theta-2\pi<\theta_{\sigma(1)}\leqq\theta_{\sigma(2)}\leqq\cdots\leqq\theta_{\sigma(n)}<\theta. \tag{7.1}$$

In the case that λ is generic, inequality holds everywhere in (7.1) and σ is uniquely determined by (λ, θ). Furthermore, σ has the following geometric interpretation: If we move along a circular arc $t-\lambda=\varepsilon e^{i\varphi}$ (with $\varepsilon>0$ and small, $\varphi\in(\theta-2\pi, \theta)$) in the positive sense, then we first cross the cut from λ to $\lambda_{\sigma(1)}$, then from λ to $\lambda_{\sigma(2)}$, etc.

If λ is *exceptional* (i.e., nongeneric), then (7.1) holds for several permutations. Letting σ denote any such (fixed) permutation, we now wish to show that there exists a system of cuts for which the above interpretation still applies (and which coincides with the system in § 6 in case λ is generic). We construct these cuts as follows.

If there is only one point λ_k on the ray $\arg(t-\lambda)=\theta_k$, we cut from λ to λ_k along the straight line segment. Otherwise, let j and l $(1\leqq j<j+l\leqq n)$ be such that (with the conventions $\theta_{\sigma(0)}=\theta-2\pi$, $\theta_{\sigma(n+1)}=\theta$)

$$\theta_{\sigma(j-1)}<\theta_{\sigma(j)}=\theta_{\sigma(j+1)}=\cdots=\theta_{\sigma(j+l)}<\theta_{\sigma(j+l+1)}.$$

Then the points $\lambda_{\sigma(j)}, \cdots, \lambda_{\sigma(j+l)}$ and no others lie on the ray $\arg(t-\lambda)=\theta_{\sigma(j)}\equiv\hat{\theta}$. We wish to cut from λ to $\lambda_{\sigma(j)}$, λ to $\lambda_{\sigma(j+1)}$, etc., by a set of nonselfintersecting polygonal arcs (meeting only at λ) such that the directed line segments constituting these arcs all have direction in $(\hat{\theta}-\varepsilon, \hat{\theta}+\varepsilon)$, where ε is any sufficiently small positive number (so that these arcs will not cross any of those from λ to λ_k for $k<j$ or $k>j+l$). First make a polygonal cut from λ to $\lambda_{\sigma(j)}$ (satisfying the above restriction) so that all points λ_k on the ray $\arg(t-\lambda)=\hat{\theta}$ which are closer to λ than $\lambda_{\sigma(j)}$ lie to the *left* of the arc. Next make a cut from λ to $\lambda_{\sigma(j+1)}$ starting out from λ slightly to the left of the previous cut and staying always to the right of all points λ_k closer to λ that $\lambda_{\sigma(j+1)}$ (except that the cut must pass to the left of $\lambda_{\sigma(j)}$ if $\lambda_{\sigma(j+1)}$ is further out than $\lambda_{\sigma(j)}$). (For the point on the ray $\arg(t-\lambda)=\hat{\theta}$ that is closest to λ we will cut along the straight line segment.) Continuing in this manner, we always begin the next cut slightly to the left of the previous one and stay to the right of all points that have not already been treated. This procedure then gives us a system of cuts so that when we traverse (in the positive sense) a circular arc close to λ with argument close to $\hat{\theta}$, we first encounter the cut from λ to $\lambda_{\sigma(j)}$, then from λ to $\lambda_{\sigma(j+1)}$, etc., and finally the one from λ to $\lambda_{\sigma(j+l)}$.

Remark 7.1. Given an admissible pair (λ, θ) and a permutation σ satisfying (7.1), we call any system of cuts of the type described above a *canonical system corresponding to the data* λ, θ, σ. (Note that in the generic case there is a unique system, whereas in the exceptional cases there are infinitely many possible systems. However, if we would call two systems *equivalent* when one can be homotopically deformed into the other without crossing any of the points $\lambda_1, \cdots, \lambda_n$ or the cut from λ to ∞, then the permutation uniquely specifies a system of cuts up to equivalence.) We also remark that given a canonical system, it would, of course, be possible to select an enumeration of $\lambda_1, \cdots, \lambda_n$ so that the cuts in $(\theta-2\pi, \theta)$ occur in their natural order, but this ordering clearly depends on the location of λ. Since we wish to discuss different choices of λ for the same $\lambda_1, \cdots, \lambda_n$, that point of view would not be convenient here.

In either case (generic or exceptional) we will use the symbol $\mathscr{P}=\mathscr{P}(\lambda, \theta, \sigma)$ for the t-plane with a specified canonical system of cuts.

Given any canonical system of cuts, one sees that the region

$$\theta_j + \pi/2 < \arg(\lambda_j - t) < \theta_j + 3\pi/2, \qquad |t - \lambda_j| < \varepsilon$$

(for sufficiently small $\varepsilon > 0$) is contained in $\mathcal{P}$ for every $j = 1, 2, \cdots, n$. Hence for such t we may define $\xi_j(s, t)$ by (3.1) and analytically continue it (with respect to t) into all of $\mathcal{P}$. Likewise, for $t \in \mathcal{P}$ we may use (6.2) to define $\Phi(s, t)$ and for the same reasons as stated there and in Remark 6.1, Proposition 6 holds for these functions as well, except for the expressions (6.4) and (6.6), which depend upon the assumption that the cuts are almost parallel (with ν determined by (6.1)). Our goal is to determine how the permutation σ influences the expressions for $C(s)$ and $\Delta(s)$ in (6.4) and (6.6) and to show in the process how to calculate all the matrices $C(s)$ and $\Delta(s)$ if they are known for a single admissible pair λ, θ.

As preparation for this, we consider the following rather general situation: Given an admissible pair (λ, θ) and a canonical system of cuts corresponding to a permutation σ, let $\mathcal{P}_{\lambda,\theta,\sigma} = \mathcal{P}$ denote the plane with these cuts and assume h and l are such that $\arg(\lambda_h - \lambda) = \arg(\lambda_l - \lambda)$ and the cut from λ to λ_l *immediately follows* the cut from λ to λ_h, i.e., $\sigma(l) = \sigma(h+1)$. Then there exists a polygonal cut π in $\mathcal{P}$ (of the type described above) from λ to λ_l that *immediately precedes* the cut from λ to λ_h such that the union of π and the original cut from λ to λ_l is a closed Jordan curve containing λ_h in its interior and *none* of the other λ_k, $k \neq h$.

Let $\tilde{\mathcal{P}}$ denote the t-plane with the same system of cuts as in $\mathcal{P}$, but the original cut from λ to λ_l is replaced by π, and let $\tilde{\xi}_j(s, t)$ and $\tilde{\varphi}_j(s, t)$ denote the corresponding analytic continuations (with respect to $t \in \tilde{\mathcal{P}}$) of the functions given locally by (3.1), (6.2). (We continue to use the notation $\xi_j(s, t)$, $\varphi_j(s, t)$ for the analytic continuations of the same germs in $\mathcal{P}$.) Let $\tilde{C}(s)$ and $\tilde{\Delta}(s)$ denote the unique matrices associated with these functions (by Proposition 6). We state the relation between these matrices and the original ones now as follows.

LEMMA 1. *In the situation described above, suppose that λ_l and λ_h are such that the cut from λ to λ_l lies on the left-hand side of the cut from λ to λ_h (case* I*). Then*

$$\tilde{C}(s) = (I - c_{hl}(s)E_{hl})C(s)(I + e^{2\pi i(\lambda_l' - s)}c_{lh}(s)E_{lh}) \tag{7.2}$$

and

$$\tilde{\Delta}(s) = (I - c_{hl}(s)E_{hl})\Delta(s), \tag{7.3}$$

where E_{hl} denotes the $n \times n$ matrix, in which all entries are zero except for a one in the (h, l) position.

If the cut from λ to λ_l lies on the right-hand side of the cut from λ to λ_h (case II*) we have*

$$\tilde{C}(s) = (I + e^{2\pi i(\lambda_l' - s)}c_{hl}(s)E_{hl})C(s)(I - c_{lh}(s)E_{lh}) \tag{7.4}$$

and

$$\tilde{\Delta}(s) = (I + e^{2\pi i(\lambda_l' - s)}c_{hl}(s)E_{hl})\,\Delta(s). \tag{7.5}$$

Proof. We give the proof only for case I; the proof in case II follows similarly. For fixed j, k, $1 \leqq j, k \leqq n$, we will choose a path in $\tilde{\mathcal{P}}$ from λ_j to λ_k along which we will analytically continue our functions. If neither j nor k equals h or else if $j = k = h$, then the path may be chosen so that it does not cross the cut from λ to λ_h (in $\mathcal{P}$); hence $\xi_j(s, t)$ and $\tilde{\xi}_j(s, t)$ (which are equal by definition for t close to λ_j and arbitrary s) remain equal as t approaches λ_k along this path. Hence $\tilde{c}_{kj}(s) = c_{kj}(s)$.

Next consider $k = h$, $j \neq h$. Then using the fact that the cut from λ to λ_h lies to the right of the cut from λ to λ_l, one sees that there exists such a path (in $\mathcal{P}$) crossing the cut from λ to λ_l exactly once and in the positive sense (i.e., the selected branch of $\arg(t-\lambda_l)$ jumps by -2π as the cut is crossed). Furthermore, we may also take the crossing point close to λ_l. Since for $s \not\equiv \lambda_l' \pmod 1$, $t \in \mathcal{P}$

$$\xi_j(s, t) = \xi_l(s, t) c_{lj}(s)(1 - e^{2\pi i(s-\lambda_l')})^{-1} + \operatorname{reg}(t-\lambda_l),$$

and since (using (3.1)) the analytic continuation of $\xi_l(s, t)$ across the cut from λ to λ_l in the positive sense is given by

$$\xi_l(s, t)\, e^{2\pi i(s-\lambda_l')},$$

then the analytic continuation of $\xi_j(s, t)$ along the path to a point close to λ_h is given by

$$\xi_j(s, t) - \xi_l(s, t) c_{lj}(s).$$

But since $\xi_j(s, t)$ and $\tilde{\xi}_j(s, t)$ have the same local definition near λ_j and the above path stays in $\tilde{\mathcal{P}}$, then the above value is by definition $\tilde{\xi}_j(s, t)$. Hence

$$\tilde{\xi}_j(s, t) = \xi_j(s, t) - \xi_l(s, t) c_{lj}(s),$$

for t close to λ_h. This expression has been shown for $s \not\equiv \lambda_l' \pmod 1$, but since all the terms on both sides are entire functions of s, it remains valid for all complex s. Using this and the relations (6.3) we obtain

$$\tilde{c}_{hj}(s) = c_{hj}(s) - c_{hl}(s) c_{lj}(s), \qquad j \neq h, \quad 1 \leqq j \leqq n. \tag{7.6}$$

Last, we consider the case $j = h$, $k \neq h$. Then there is a path (in $\mathcal{P}$) from λ_h to λ_k that crosses the cut from λ to λ_l exactly once and in the negative sense (also crossing close to λ_l). Using the same reasoning as above, the analytic continuation of $\xi_k(s, t)$ along this path to a point near λ_k is given by

$$\xi_h(s, t) + \xi_l(s, t) c_{lh}(s)\, e^{2\pi i(\lambda_l' - s)},$$

and comparing this with the analytic continuation in $\tilde{\mathcal{P}}$ we find

$$\tilde{\xi}_h(s, t) = \xi_h(s, t) + \xi_l(s, t) c_{lh}(s)\, e^{2\pi i(\lambda_l' - s)}.$$

Using this and the relations (6.3) we obtain

$$\tilde{c}_{kh}(s) = c_{kh}(s) + c_{kl}(s) c_{lh}(s)\, e^{2\pi i(\lambda_l' - s)}, \qquad k \neq h, \quad 1 \leqq k \leqq n. \tag{7.7}$$

From (7.6), (7.7) and the definition $c_{jj} = 1 - e^{2\pi i(s-\lambda_j')}$, we obtain (7.2).

In a similar manner, observe that for every j, $1 \leqq j \leqq n$, $\tilde{\varphi}_j(s, t) = \varphi_j(s, t)$ for t close to λ_k, $k \neq h$, $1 \leqq k \leqq n$ (since φ_j, $\tilde{\varphi}_j$ have the same definition for $|t|$ large, $\arg t = \theta - \pi$, and all points λ_j for $j \neq h$ are accessible from such points in both $\mathcal{P}$ and $\tilde{\mathcal{P}}$). But for t close to λ_h, we have (using (6.5))

$$\tilde{\varphi}_j(s, t) = \varphi_j(s, t) - \xi_l(s, t)\, \delta_{lj}(s).$$

This implies (for $j = 1, 2, \cdots, n$) that

$$\tilde{\delta}_{kj}(s) = \delta_{kj}(s), \qquad k \neq h, \quad 1 \leqq k \leqq n,$$

and

$$\tilde{\delta}_{hj}(s) = \delta_{hj}(s) - c_{hl}(s)\, \delta_{lj}(s),$$

which yields (7.3). This completes the proof of case I and the lemma.

We now return to the question of how the permutation σ affects the matrices $C(s)$ and $\Delta(s)$. Corresponding to a fixed, but arbitrary, canonical system of cuts, let σ denote the permutation. We define two sets σ^+ and σ^- of ordered pairs (j, k) according to the rule

$$(7.8) \qquad \begin{aligned} &(j, k) \in \sigma^+ \quad \text{iff } \sigma^{-1}(j) > \sigma^{-1}(k), \\ &(j, k) \in \sigma^- \quad \text{iff } (k, j) \in \sigma^+. \end{aligned}$$

One may think of σ^+ as defining a complete ordering of the numbers $1, 2, \cdots, n$; in case of almost parallel cuts, the ordering coincides with the dominance relation in $S'_{\nu+1} = S(\tau_\nu, \tau_{\nu+1})$ for ν as in (6.1). (Compare [1, § 3.1] for the notation and interpretation of the dominance relation in terms of the ordering of the cuts.) The purpose of σ^+ and σ^- is to provide (in the general case) the structure that allows $C(s)$ to be decomposed in a manner similar to (6.4). We state our main result in this section as follows.

THEOREM 3. *Let a differential equation* (0.1), *a formal fundamental solution matrix* $H(z)$, *a solution* $X(z)$ *in Floquet form, an admissible pair* (λ, θ) *and a canonical system of cuts be given. (Consider all of the above fixed, but arbitrary, subject to our assumptions and natural restrictions.) If* σ *denotes the permutation induced by the cuts, let* $\sigma^\pm$ *be defined by* (7.8). *Then Proposition* 6 *extends to this situation as follows.*

(i) *There exist unique constant (i.e., independent of* s*) matrices* $C^+ = C^+[\sigma]$, *resp.* $C^- = C^-[\sigma]$ *having ones along their diagonals and nonzero off-diagonal elements only in positions* $(j, k) \in \sigma^+$, *resp.* $(j, k) \in \sigma^-$ *and such that*

$$(7.9) \qquad C(s) = C^- - e^{2\pi i(sI - \Lambda')} C^+.$$

(ii) *There is a unique constant invertible matrix* $\Omega = \Omega[\sigma]$ *such that*

$$(7.10) \qquad \Delta(s) = \Omega(I - e^{2\pi i(sI - M)}) \quad \textit{with } M \textit{ as in } (0.3).$$

(iii) *If we define*

$$(7.11) \qquad e^{2\pi i M[\sigma]} = C^-[\sigma](C^+[\sigma])^{-1} e^{2\pi i \Lambda'},$$

then

$$(7.12) \qquad e^{2\pi i M[\sigma]} = \Omega[\sigma] e^{2\pi i M} \Omega^{-1}[\sigma].$$

Proof. First observe that slight changes in θ do not affect the validity of statements (a) and (b) in Proposition 6 nor the form of $C(s)$ or $\Delta(s)$. Hence we may assume without loss of generality that $\theta \neq \eta_\nu$ (for every integer ν) and at the same time that none of the points $\lambda_1, \cdots, \lambda_n$ lie on the straight line through λ with direction θ. (This line we denote by g.) Consider such a straight line g now to be fixed and think of varying λ along g, observing that all pairs (λ, θ) are admissible for all $\lambda \in g$. For λ sufficiently far out (and on the proper side of g), the cuts can be seen to be almost parallel and in this situation it follows from Proposition 6 that statement (i) holds if we take

$$C^- = V_\nu \cdots V_{\nu-\mu+1} \quad \text{and} \quad C^+ = (V_{\nu-\mu} \cdots V_{\nu-m+1})^{-1}.$$

(Compare with [1, § 3], to see that $C^\pm$ have the required form.) Moreover, using (6.6) and (1.6) we see that (ii) and (iii) hold with $\Omega[\sigma] = \Omega_\nu$ and $M[\sigma] = M_\nu$ since

$$C^-(C^+)^{-1} e^{2\pi i \Lambda'} = e^{2\pi i \Lambda'} V_{\nu+m} \cdots V_{\nu+1} = e^{2\pi i M_\nu}$$

(using that for every integer l, $V_{l+m} = e^{-2\pi i \Lambda'} V_l e^{2\pi i \Lambda'}$).

Note that as λ moves along g in an interval containing only *generic values*, then using (6.3) and (6.5) it is easy to see that $c_{kj}(s)$ and $\delta_{kj}(s)$ do not change. Because of the choice of g, the only nongeneric points on g are isolated and there are only finitely many of them. So to establish the validity of (7.9), (7.10) and (7.12) it suffices to consider what changes take place in $C(s)$, $\Delta(s)$ for generic values of λ and $\tilde{\lambda}$ in adjacent open intervals that are separated by an exceptional value $\hat{\lambda}$.

Let the permutations corresponding to the cuts from λ, resp., $\tilde{\lambda}$, be denoted by σ, resp. $\tilde{\sigma}$, and note that since $\hat{\lambda}$ is the only exceptional value between λ and $\tilde{\lambda}$, then both

$$(\hat{\lambda}, \theta, \sigma) \quad \text{and} \quad (\hat{\lambda}, \theta, \tilde{\sigma})$$

are admissible data (in the sense that they correspond to systems of cuts of the type discussed above). Also note that with respect to the points λ_h, λ_l with $\arg(\lambda_h - \lambda) = \arg(\lambda_l - \lambda)$, the difference between σ and $\tilde{\sigma}$ corresponds to a reversal of the ordering of the cuts. There are only finitely many such rays emanating from $\hat{\lambda}$ containing points which are thus affected and the permutation taking σ to $\tilde{\sigma}$ acts separately on the cuts corresponding to each of the rays. Observe that in the limiting situation as $\lambda \to \hat{\lambda}$, the elements in $C(s)$ and $\Delta(s)$ depend only on the permutation σ and not on the specific cuts used in the limiting situation $(\hat{\lambda}, \theta, \sigma)$, since all points $\lambda_1, \cdots, \lambda_n$ are accessible to each other and λ by paths lying in both planes. To now see what changes take place between $(\hat{\lambda}, \theta, \sigma)$ and $(\hat{\lambda}, \theta, \tilde{\sigma})$, observe that there exists a sequence of permutations taking σ into $\tilde{\sigma}$ corresponding to systems of cuts such that for two consecutive systems of cuts the situation of Lemma 1 applies. Hence to show that the conclusions of Theorem 3 hold for $(\hat{\lambda}, \theta, \tilde{\sigma})$ (and consequently for $(\tilde{\lambda}, \theta, \tilde{\sigma})$) it is sufficient to prove the following.

LEMMA 2. *In the situation of Lemma* 1, *let* σ, *resp.* $\tilde{\sigma}$, *denote the permutations corresponding to* $\mathcal{P}$, *resp.*, $\tilde{\mathcal{P}}$, *satisfying the aforementioned conditions and let* $\sigma^{\pm}$, $\tilde{\sigma}^{\pm}$ *be determined by* (7.8). *Furthermore, assume that the conclusions of Theorem* 3 *hold for* $\hat{\lambda}$, θ, σ (*with matrices denoted* $C(s)$, $C^{\pm}$, $\Delta(s)$, Ω). *Then for* $\hat{\lambda}$, θ, $\tilde{\sigma}$ *we have*

$$\tilde{C}(s) = \tilde{C}^{-} - e^{2\pi i(sI - \Lambda')}\tilde{C}^{+}, \tag{7.13}$$

$$\tilde{\Delta}(s) = \tilde{\Omega}(I - e^{2\pi i(sI - M)}), \tag{7.14}$$

and

$$\tilde{\Omega}\, e^{2\pi i M}\tilde{\Omega}^{-1} = \tilde{C}^{-}(\tilde{C}^{+})^{-1}\, e^{2\pi i \Lambda'}, \tag{7.15}$$

where (*in case* $(h, l) \in \sigma^{-}$) *we have*

$$\tilde{C}^{+} = (I - c_{hl}^{-}\, e^{2\pi i(\lambda_h' - \lambda_l')}E_{hl})C^{+}(I - c_{lh}^{+}E_{lh}), \tag{7.16}$$

$$\tilde{C}^{-} = (I - c_{hl}^{-}E_{hl})C^{-}(I - c_{lh}^{+}E_{lh}), \tag{7.17}$$

and

$$\tilde{\Omega} = (I - c_{hl}^{-}E_{hl})\Omega, \tag{7.18}$$

whereas (*in case* $(h, l) \in \sigma^{+}$) *we have*

$$\tilde{C}^{+} = (I - c_{hl}^{+}E_{hl})C^{+}(I - c_{lh}^{-}E_{lh}), \tag{7.19}$$

$$\tilde{C}^{-} = (I - e^{2\pi i(\lambda_l' - \lambda_h')}c_{hl}^{+}E_{hl})C^{-}(I - c_{lh}^{-}E_{lh}), \tag{7.20}$$

and

$$\tilde{\Omega} = (I - e^{2\pi i(\lambda_l' - \lambda_h')}c_{hl}^{+}E_{hl})\Omega. \tag{7.21}$$

Hence $\tilde{C}^{+}$, *resp.* $\tilde{C}^{-}$ *have ones on their main diagonals and nonzero off-diagonal elements only for positions in* $\tilde{\sigma}^{+}$, *resp.* $\tilde{\sigma}^{-}$.

Proof. We will restrict ourselves to the case $(h, l) \in \sigma^-$; the other case follows analogously. Observe that $(h, l) \in \sigma^-$ corresponds to case I of Lemma 1, since $\sigma^{-1}(l) > \sigma^{-1}(h)$ means that the cut from λ to λ_l comes after the cut from λ to λ_h (when traversing an arc close to λ in the positive sense). From (7.9) we find $c_{hl}(s) = c_{hl}^-$, $c_{lh}(s) = -e^{2\pi i(s-\lambda_l)} c_{lh}^+$. From (7.2) we therefore find (7.13) (if we define $\tilde{C}^{\pm}$ by (7.16), (7.17). Moreover, (7.14) and (7.15) follow immediately, using (7.3), (7.10), (7.16), (7.17), (7.18). It remains to show that $\tilde{C}^+$, $\tilde{C}^-$ (defined by (7.16), (7.17)) again have ones along the diagonal and nonzero off-diagonal terms only in positions $(j, k) \in \tilde{\sigma}^+$ resp. $(j, k) \in \tilde{\sigma}^-$. In order to do so, observe that re-enumerating the points $\lambda_1, \cdots, \lambda_n$ corresponds to applying the same permutation similarity transform to $C(s)$ and $\tilde{C}(s)$ (hence to $C^{\pm}$ and $\tilde{C}^{\pm}$). So for the purpose of simplifying the arguments, we may assume (without loss of generality) that $\lambda_1, \cdots, \lambda_n$ are enumerated according to the ordering of the cuts in $\mathcal{P}$ (from right to left when looking toward λ). In this case $(j, k) \in \sigma^+$ iff $j < k$, i.e., C^+ is upper triangular and C^- lower triangular. Moreover, the selection of cuts and the fact that we consider the case $(h, l) \in \sigma^-$ implies $h = l+1$. Hence it follows easily from (7.16), (7.17) that $\tilde{C}^+$ (resp. $\tilde{C}^-$) has ones along the diagonal and a zero in the (l, h) (resp. (h, l)) position, and the only nonzero element below (above) the diagonal may occur in the (h, l) (resp. (l, h)) position. This completes the proof.

8. Applications and final remarks. Finally, we will discuss how the previous results may be used to solve the central connection problem, in the case of an arbitrary admissible (λ, θ) (λ generic). For a solution $X(z) = L(z) z^M$ in Floquet form with monodromy matrix M in upper triangular Jordan form, let $\Delta(s) = \Omega(I - e^{2\pi i(sI - M)})$ be as in Theorem 3. If $b_j(p)$, μ, n_1 are as in Theorem 2, then the proof of that theorem also applies to situations of arbitrary generic λ (where the cuts generally are not almost parallel), and we find for arbitrary p and $j = 1, \cdots, n_1$:

$$b_j(p) = \sum_{q=0}^{j-1} \frac{1}{q!} \sum_{k=1}^{n} \omega_{k, j-q} \left(\frac{\partial}{\partial s}\right)^q \xi_k(s, \lambda) \Big|_{s=p+\mu+1}. \tag{8.1}$$

This expression and the corresponding formulas for the remaining blocks of M allow the following interpretations:

(a) If a monodromy matrix M in Jordan form and one Laurent coefficient $B(p) = [b_1(p), \cdots, b_n(p)]$ of a corresponding solution in Floquet form are known, and if the matrix $\xi(s, \lambda)$ is invertible (at least) for $s = p + \mu_j + 1$, $1 \leqq j \leqq n$, then the matrix $\Omega = [\omega_{kj}]$ can be uniquely computed from (8.1) and the corresponding formulas for the remaining blocks of M. Moreover, from (7.12), (7.11), (7.9) we can, successively, compute C^-, C^+, $C(s)$, and with the help of Lemma 1 we can explicitly see how $C(s)$ (and $\Delta(s)$) changes with respect to λ so that in the end we may compute the corresponding matrices for a situation of almost parallel cuts. Using Proposition 6 we can then compute all the Stokes' multipliers and all the central connection matrices. We wish to emphasize that all the computations required above are merely matrix algebraic operations, except for the evaluation of $\xi(s, t)$ (and its s-derivatives) at the point $t = \lambda$. In case the so-called "pentagonal condition" holds, i.e.,

$$|\lambda_j - \lambda| < |\lambda_j - \lambda_k| \qquad (k \neq j, 1 \leqq j, k \leqq n),$$

then for every $j = 1, \cdots, n$, the series (3.1) converges at $t = \lambda$ and may therefore be used directly to evaluate $\xi(s, \lambda)$. Generally one has to analytically continue $\xi(s, t)$ (along a straight line) to the point $t = \lambda$. This can be done by explicit summability methods or by rewriting (3.1) as a convergent generalized factorial series in s. (See R.

Schäfke [17] whose proof, given only for an equation (0.4), also carries over to equations of the form (0.1).) Concerning the invertibility of $\xi(s, \lambda)$, we may use either the asymptotic obtained in (3.3) to see that $\xi^{-1}(s, \lambda)$ exists for Re s sufficiently large, or, in the case of an equation (0.4) we can explicitly compute the determinant of $\xi(s, t)$ (using both the difference and the differential equation for $\xi(s, t)$ and its asymptotic; compare Hukuhara [9] for details):

$$\det \xi(s, t) = \prod_1^n \frac{(\lambda_j - t)^{s-\lambda_j'-1}}{\Gamma(s-\mu_j)}. \tag{8.2}$$

Consequently, if we assume that A_1 has n eigenvalues $\mu_1, \cdots, \mu_n$ incongruent modulo one, then $\xi(s, \lambda)$ is invertible for the values $s = \mu_j + 1$ $(1 \leqq j \leqq n)$. Hence if $X(z) = L(z)z^M$ (with $M = \operatorname{diag}[\mu_1, \cdots, \mu_n]$, $L(z)\, e^{-\lambda z} = \sum_0^\infty B(p)z^p$) is a given solution in Floquet form, then (8.1) (and the corresponding formulas for the other—one-dimensional—Jordan blocks of M) imply (with $p = 0$)

$$b_j(0) = \sum_{k=1}^n \omega_{k,j}\xi_k(\mu_j + 1, \lambda) \qquad (1 \leqq j \leqq n),$$

and solving these systems of linear equations gives Ω explicitly. Kohno [14], under the additional assumption of a pentagonal condition (and considering only the case $\lambda = 0$), derived an analogous result (even for equations having a singularity of larger rank at ∞ and a somewhat more general structure for the formal solution, but zero, the only finite singularity, being of first kind). To remove the assumptions of a pentagonal condition from Kohno's results, it has been observed by Sibuya [24] that one should in some sense analytically continue the functions in Kohno's formulas. One natural way to do this involves continuing the functions $\xi_j(s, t)$ with respect to t to a convenient value λ and then one needs to know how the connection constants depend upon the choice of λ.

(b) Using the freedom in the selection of a solution in Floquet form (corresponding to some fixed M), the result (a) has the following "converse": Suppose that the Stokes' multipliers are known, then one can compute the matrix $C(s)$ (and vice versa) corresponding to our selected admissible pair (λ, θ). From (7.9) we then compute C^+, C^-. If Ω is *any* invertible constant matrix for which

$$\Omega^{-1}C^-(C^+)^{-1}\, e^{2\pi i\Lambda'}\Omega = e^{2\pi iM}$$

with M in upper triangular Jordan form, then by (8.1) and the corresponding formulas for the other blocks of M we *define* $B(p) = [b_1(p), \cdots, b_n(p)]$, for arbitrary integer p. The matrix

$$X(z) = L(z)z^M, \qquad L(z)\, e^{-z\lambda} = \sum_{-\infty}^{\infty} B(p)z^p$$

then is a fundamental solution of (0.1) in Floquet form whose central connection coefficients can be explicitly computed in terms of Ω (following the same steps as in (a)).

As a result of the preceding discussion, we now formulate Theorem 4.

THEOREM 4. *Consider a differential equation* (0.1) *satisfying our basic assumptions, any selected formal fundamental solution matrix* $H(z)$ *as in* (0.2), *and any admissible pair* (λ, θ), with λ *a generic value. Then the following statements hold*:

(a) *If a monodromy matrix* M (*in upper triangular Jordan canonical form*) *and the* p*th Laurent coefficient* $B(p)$ *of* $L(z)\, e^{-z\lambda}$ (*for a solution* $X(z) = L(z)z^M$ *in Floquet form*) *are known* (*with fixed* p, *sufficiently large*), *then the matrix* $\Omega = \Delta(s)\ (I - e^{2\pi i(sI-M)})^{-1}$ *can be explicitly computed from systems of linear equations involving* $B(p)$ *and* $\xi(s, \lambda)$

together with its derivatives with respect to s, for values $s = p + \mu_j + 1$ $(1 \leqq j \leqq n)$. *Moreover, from* Ω *one can explicitly compute all the Stokes' multipliers and central connection matrices, using only arithmetical operations.*

(*b*) *If all the Stokes' multipliers (or equivalently,* $C(s)$*) have been calculated from the global behavior of* $\xi(s, t)$, *and if* Ω *is any invertible matrix for which*

$$\Omega^{-1} C^{-} (C^{+})^{-1} e^{2\pi i \Lambda'} \Omega = e^{2\pi i M}$$

with M in upper triangular Jordan form, then defining $B(p)$ *by* (8.1) *resp. the corresponding formulas for the other blocks of M, the matrix*

$$X(z) = L(z) z^{M}, \qquad L(z)\, e^{-z\lambda} = \sum_{-\infty}^{\infty} B(p) z^{p},$$

is a fundamental solution of (0.1) *in Floquet form, and its central connection coefficients may be explicitly computed from* Ω, *using the same procedure as in* (a).

9. An example. Consider the scalar, second-order differential equation

$$\frac{d^2y}{dz^2} + \frac{1}{z}\frac{dy}{dz} + \left(h^2 - \frac{\tilde{\lambda}}{z^2} + \frac{h^2}{z^4}\right) y = 0, \tag{9.1}$$

which we write in equivalent system form as

$$\frac{d}{dz}\begin{pmatrix} y \\ dy/dz \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -h^2 + \tilde{\lambda}/z^2 - h^2/z^4 & -1/z \end{pmatrix} \begin{pmatrix} y \\ dy/dz \end{pmatrix} \equiv A(z) \begin{pmatrix} y \\ dy/dz \end{pmatrix}. \tag{9.2}$$

It is well known that (9.1) is related to Mathieu's differential equation

$$\frac{d^2\tilde{y}}{d\theta^2} + (\tilde{\lambda} - 2h^2 \cos 2\theta)\tilde{y} = 0 \tag{9.3}$$

through the change of variable $z = e^{i\theta}$ and $y(z) = \tilde{y}(\theta)$. The quantities h and $\tilde{\lambda}$ are complex parameters with $h \neq 0$: We use the same notation as in Meixner and Schäfke [22] except we have been forced to replace their λ by $\tilde{\lambda}$, since λ has already a special meaning for us from §§ 6–8.

Traditionally, it has been more popular [22] to treat (9.3) using Floquet's theory for periodic differential equations rather than (9.1) or (9.2), which are meromorphic differential equations having irregular singularities at 0 and ∞. Our goal is to apply the preceding theory to represent a fundamental matrix for (9.2) in Floquet form, that is, as a convergent Laurent series times a matrix power z^M. In terms of (9.3) this corresponds to calculating the Fourier expansion for the periodic part of a Floquet solution as well as the Floquet exponents. The formulas we derive here are not intended to supplant or duplicate the rather extensive literature on Mathieu functions, much of which is motivated by applications to eigenfunction expansions, but instead to indicate how the results of this paper could be applied to an interesting differential equation with multiple irregular singularities.

Near ∞, we select for (9.2) the formal fundamental solution matrix

$$H(z) = F(z) z^{\Lambda'} e^{\Lambda z}, \tag{9.4}$$

where $\Lambda = \operatorname{diag}\{ih, -ih\}$, $\Lambda' = \operatorname{diag}\{-\frac{1}{2}, -\frac{1}{2}\}$, $F(z) = \sum_0^{\infty} F_p z^{-p}$ with

$$F_0 = \begin{pmatrix} 1 & 1 \\ ih & -ih \end{pmatrix},$$

and h is selected so that $-\pi < \arg h \leqq 0$, replacing h by $-h$ if necessary. This selection corresponds to ordering the pair $(ih, -ih)$ lexicographically (in decreasing order, first by real, then imaginary parts).

In the study of (9.3) it is convenient to use its special form and periodicity to minimize the required calculations. For (9.1) this corresponds to its invariance with respect to each of the substitutions $z \to z e^{\pi i}$ and $z \to 1/z$. For (9.2) this implies that if $X(z)$ is a fundamental solution matrix (actual or formal), then

$$\operatorname{diag}\{1, -1\}X(z e^{\pi i}) \quad \text{and} \quad \operatorname{diag}\{1, -1/z^2\}X(1/z)$$

also are fundamental solutions. (We remark that all solutions are considered to be defined on the Riemann surface of $\log z$ with $\arg z$ selected consistent with the treatment in the earlier sections.)

It follows that there exists a constant, invertible matrix C such that

$$\operatorname{diag}\{1, -1\}H(z e^{\pi i}) = H(z)C \tag{9.5}$$

and using (9.4) we find that

$$C = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}$$

and

$$\operatorname{diag}\{1, -1\}F(-z) = F(z)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{9.6}$$

Also, since

$$\begin{aligned} \hat{H}(z) &= \operatorname{diag}\{1, -1/z^2\}H(1/z) \\ &= \operatorname{diag}\{1, -1/z^2\}F(1/z)z^{-\Lambda'} e^{\Lambda z^{-1}} \end{aligned}$$

is a formal fundamental solution, we may select this $\hat{H}(z)$ as our formal fundamental solution near $z = 0$.

The Stokes' rays for (9.2) with respect to the singularity at ∞ are those rays $\arg z = \tau$ for which $\operatorname{Re}(2ihz)$ changes sign. We let

$$\tau_0 = -\arg h \quad \text{and define } \tau_\nu = \tau_0 + \nu\pi$$

for all integers ν. Let $X_\nu(z)$ denote the νth normal solution (with respect to the singularity at ∞ and our selected $(H(z))$ and let (V_ν) denote the corresponding system of Stokes' multipliers.[1] From

$$X_\nu(z) \cong H(z), \quad z \to \infty, \quad \arg z \in (\tau_{\nu-1}, \tau_{\nu+1}),$$

we have (using (9.5))

$$\operatorname{diag}\{1, -1\}X_\nu(z e^{\pi i})C^{-1} \cong H(z), \quad z \to \infty, \quad \arg z \in (\tau_{\nu-2}, \tau_\nu),$$

hence (by uniqueness of the normal solutions)

$$X_{\nu-1}(z) = \operatorname{diag}\{1, -1\}X_\nu(z e^{\pi i})C^{-1} \quad \text{and} \quad V_{\nu-1} = CV_\nu C^{-1} \tag{9.7}$$

for all ν, i.e.,

$$V_{\nu-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} V_\nu \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{9.8}$$

[1] Normal solutions can be represented, for example, as Laplace integrals using results of A. Erdélyi [19], [20].

In particular, due to the ordering of $(ih, -ih)$, we have

$$V_0 = \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} \quad \text{and} \quad V_1 = \begin{pmatrix} 1 & v \\ 0 & 1 \end{pmatrix},$$

where v is a complex constant.

Define $\hat{X}_\nu(z) = \operatorname{diag}\{1, -1/z^2\} X_\nu(1/z)$, where we select $\arg(1/z) = -\arg z$. Then the Stokes' rays with respect to the singularity at $z = 0$ are those rays for which $\operatorname{Re}(2ihz^{-1})$ changes sign. To simplify the notation we can enumerate them as $\hat{\tau}_\nu = -\tau_\nu$ and note that $\arg z \in (\tau_{\nu-1}, \tau_{\nu+1})$ iff $\arg(1/z) \in (\hat{\tau}_{\nu-1}, \hat{\tau}_{\nu+1})$. It follows that $\hat{X}_\nu(z)$ is the νth normal solution (at 0 with respect to the selected $\hat{H}(z)$ and the Stokes' multipliers at 0 are exactly the same as the V_ν).

Hence there is just one quantity v that determines all the Stokes' multipliers at 0 and ∞; to calculate v we apply (5.7), which in this case reduces to

$$f_1(p) = v\alpha_2(\lambda_1' - p + 1) \tag{9.9}$$

with $\alpha_2(\lambda_1' - p + 1) \cong \Gamma(p)(2ih)^{-p}(f_2(0)/(2\pi i) + o(1))$ as $p \to +\infty$.[2] In using Corollary 1 we take $\eta = \eta_1 = \arg(ih) \in (-\pi/2, \pi/2]$, $\eta_0 = \eta_1 + \pi$, and also assume that $\lambda_1' \not\equiv \mu_j \pmod 1$.

Recalling our selection for F_0, this implies

$$v = -\frac{2\pi}{h} \lim_{p \to +\infty} \frac{f_{21}(p)(2ih)^p}{\Gamma(p)}, \tag{9.10}$$

where $f_1(p) = (f_{11}(p), f_{21}(p))^T$. We remark also that from the results in § 2 as well as in [11], (9.10) holds without the extra assumption on λ_1', while the construction of the function $\alpha_2(\cdot)$ and (9.9) depend upon that assumption. (The case when $\lambda_1' \equiv \mu_j \pmod 1$ will be seen shortly to be an especially simple one.)

The coefficients $f_1(p)$ are easily seen to satisfy

$$(2ihp) f_{11}(p) = ((p - 1/2)^2 - \tilde{\lambda}) f_{11}(p-1) + h^2 f_{11}(p-3), \qquad p \geqq 1 \tag{9.11}$$

and

$$f_{21}(p+1) = -(p + 1/2) f_{11}(p) + ih f_{11}(p+1), \qquad p \geqq -1 \tag{9.12}$$

with $f_{11}(-2) = f_{11}(-1) = 0$ and $f_{11}(0) = 1$. Defining $d(p) = (2ih)^p f_{11}(p)$, $p \geqq -2$, one sees that $(p+1)\, d(p+1) = (p + \frac{1}{2} + \beta)\,(p + \frac{1}{2} - \beta)\, d(p) - 4h^4\, d(p-2)$, $p \geqq 0$, where $\beta^2 \equiv \tilde{\lambda}$. Also defining

$$c(p) = d(p) \frac{\Gamma(p+1)}{\Gamma(p + \frac{1}{2} + \beta)\Gamma(p + \frac{1}{2} - \beta)} \quad \text{for } p \geqq 0,$$

one sees that

$$c(p+1) = c(p) - \frac{4h^4 \Gamma(p+1)\, d(p-2)}{\Gamma(p + \frac{3}{2} + \beta)\Gamma(p + \frac{3}{2} - \beta)}, \qquad p \geqq 0,$$

which implies that

$$c(p) = c(0) - 4h^4 \sum_{j=0}^{p-1} \frac{\Gamma(j+1)\, d(j-2)}{\Gamma(j + \frac{3}{2} + \beta)\Gamma(j + \frac{3}{2} - \beta)}, \qquad p \geqq 0.$$

Since $\Gamma(p + \frac{1}{2} + \beta)\Gamma(p + \frac{1}{2} - \beta)/\Gamma(p+1) \cong \Gamma(p)$ as $p \to +\infty$, we conclude from (9.10) that

$$v = 2\pi i \lim_{p \to +\infty} \frac{f_{11}(p)(2ih)^p}{\Gamma(p)} = 2\pi i \lim_{p \to +\infty} c(p),$$

[2] Also see [21] where the case of a scalar, second-order differential equation is described in detail.

where we have used that the first limit exists also from (9.9). (Also see [21], where the matrix formulas have been translated to the scalar case.) Hence it follows that

$$\frac{v}{2\pi i}=\frac{1}{\Gamma(\frac{1}{2}+\beta)\Gamma(\frac{1}{2}-\beta)}-4h^4\sum_{j=0}^{\infty}\frac{\Gamma(j+3)\,d(j)}{\Gamma(j+\frac{7}{2}+\beta)\Gamma(j+\frac{7}{2}-\beta)}$$

and since the $d(j)$ are easily seen to be polynomials in $\beta^2=\tilde{\lambda}$ and h^4 and of order $O(\Gamma(j))$ as $j\to\infty$, then the above series can be shown to converge compactly with respect to both variables, hence to an entire function in $\tilde{\lambda}$ and h^4. Note that the Stokes' factors are not defined for $h=0$, but the function v has a removable singularity there, in any case.

In terms of the Stokes' multipliers, one can construct a circuit factor for $X_{-1}(z)$ as $-V_1V_0$, hence the most general circuit factor is given by

$$\Omega^{-1}(-V_1V_0)\Omega=\Omega^{-1}\begin{pmatrix}-1-v^2 & -v\\ -v & -1\end{pmatrix}\Omega\equiv e^{2\pi iM}, \tag{9.13}$$

where Ω is an invertible matrix. To put $e^{2\pi iM}$ in Jordan form, use an auxiliary parameter μ defined (up to sign and modulo 2) by the equation

$$2i\cos\pi\mu=v.$$

For every possible fixed selection of μ, the values $x_1=e^{2\pi i\mu}$, $x_2=e^{-2\pi i\mu}$ then are the roots of

$$x^2+(2+v^2)x+1=0 \tag{9.14}$$

(hence are the eigenvalues of $e^{2\pi iM}$). The roots are equal in the cases $v=0$, resp. $v^2=-4$, in which cases they are -1, resp. 1. Both special cases are of particular interest: In case $v=0$, i.e., 2μ an odd integer, both columns of $F(z)$ converge, and a solution in Floquet form is given by $H(z)$. In case $v^2=-4$, i.e., μ an integer, a Jordan canonical form for $e^{2\pi iM}$ is

$$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix},$$

and M can be selected as

$$M=\begin{pmatrix}0 & (2\pi i)^{-1}\\ 0 & 0\end{pmatrix}.$$

When $v^2\neq -4$, we select Ω as

$$\Omega=\begin{pmatrix}e^{i\pi\mu} & e^{-i\pi\mu}\\ -i & -i\end{pmatrix}, \tag{9.15}$$

and we may take $M=\operatorname{diag}\{\mu,-\mu\}$ in (9.13); but every selection $M=\operatorname{diag}\{\mu_1,\mu_2\}$, with $e^{2\pi i\mu_j}=x_j$, $j=1,2$, will also satisfy (9.13). When $v^2=-4$ we may select

$$\Omega=\begin{pmatrix}1 & 0\\ \mp i & \pm i/2\end{pmatrix}\quad\text{in case } v=\pm 2i. \tag{9.16}$$

To calculate the Laurent coefficients, we first construct the associated functions

$$\xi_1(s,t)=\sum_0^{\infty}\frac{f_1(l)(ih-t)^{l+s-1/2}}{\Gamma(l+s+\frac{1}{2})}\quad\text{and}\quad\xi_2(s,t)=\sum_0^{\infty}\frac{f_2(l)(-ih-t)^{l+s-1/2}}{\Gamma(l+s+\frac{1}{2})},$$

where (consistent with the discussion in § 6) we may choose $\lambda = 0$ and cut the complex t-plane from $\pm ih$ to 0 along the line segments and from 0 to ∞ along $\arg(ih) + 3\pi/2 \equiv \theta$, with $\theta_1 = \arg(ih) \in (-\pi/2, \pi/2]$.

For t in the cut plane, denoted $\mathcal{P}$, we select

$$\arg(ih - t) \in (\theta_1, \theta_1 + 2\pi) \quad \text{and} \quad \arg(-ih - t) \in (\theta_1 + \pi, \theta_1 + 3\pi)$$

(for t close to ih, resp. $-ih$) to define the nonintegral powers. The identity (9.6) has the following interpretation for the associated functions.

Replacing t by $-t$ in the definition of $\xi_1(s, t)$ and using (9.6), we have

$$(ih + t)^{1/2-s}\xi_1(s, -t) = \operatorname{diag}\{1, -1\}(-ih - t)^{1/2-s}\xi_2(s, t) \tag{9.17}$$

for t such that $t, -t \in \mathcal{P}$. For $t \in \mathcal{P}$ with $\operatorname{Re}(te^{-i\theta_1}) < 0$ we find $\arg(-ih - t) = \pi + \arg(ih + t)$; hence

$$\xi_1(s, -t) = e^{i\pi(1/2-s)} \operatorname{diag}\{1, -1\}\xi_2(s, t).$$

Using the above relations in

$$\xi_1(s, t) = c_{21}(s)(1 - e^{2\pi i(s-1/2)})^{-1}\xi_2(s, t) + \operatorname{reg}(t + ih)$$

with $t \to -t$, one sees that

$$c_{12}(s) = c_{21}(s)\, e^{2\pi i(s-1/2)} = -c_{21}(s)\, e^{2\pi i s},$$

which agrees with (6.4) when we take $\nu = -2$ in (6.1).

Note that the above choice of $\lambda = 0$ is generic (in the sense of § 6) but the system of cuts is *not* almost parallel. This choice for λ is especially convenient, however, because both power series converge absolutely there (like geometric series) and the Laurent coefficients are produced directly rather than multiplied by the power series for $e^{-\lambda z}$ if $\lambda \neq 0$.

To apply formula (8.1), let μ_1, μ_2 be defined as above and let

$$\xi_k(p + \mu_j + 1, 0) = \sum_{l=0}^{\infty} \frac{f_k(l)[(-1)^{k+1}(ih)]^{l+p+\mu_j+1/2}}{\Gamma(l + p + \mu_j + 1)}$$

for $j, k = 1, 2$ and $p \in \mathbb{Z}$. (Note that in defining the nonintegral powers above, we must select the "forbidden" values $\arg(ih) = \theta_1$ and $\arg(-ih) = \pi + \theta_1$ according to the discussion preceding the statement of Theorem 2.)

In case $v^2 \neq -4$, take $\Omega = (\omega_{kj})$ as in (9.15) and form

$$l_j(p) \equiv b_j(p, 0) = \omega_{1j}\xi_1(p + \mu_j + 1, 0) + \omega_{2j}\xi_2(p + \mu_j + 1, 0) \quad \text{for } j = 1, 2. \tag{9.18}$$

Then from Theorem 4(b), $\sum_{-\infty}^{+\infty} [l_1(p), l_2(p)]z^{pI+M}$ *is a fundamental solution matrix for* (9.2) *in Floquet form.*

If $v^2 = -4$, then $\mu_1 = 0$ is a double root and we calculate $\xi_k(p + 1, 0)$, $k = 1, 2$, just as above and also

$$\frac{\partial}{\partial s}\xi_k(s, 0)\bigg|_{s=p+1} = \sum_{l=0}^{\infty} \frac{f_k(l)[(-1)^{k+1}ih]^{l+p+1/2}}{\Gamma(l + p + \frac{3}{2})} \log((-1)^{k+1}ih)$$

$$- \sum_{l=0}^{\infty} \frac{f_k(l)[(-1)^{k+1}ih]^{l+p+1/2}}{(\Gamma(l + p + \frac{3}{2}))^2}\Gamma'(l + p + \tfrac{3}{2}), \qquad k = 1, 2.$$

Then using Ω as in (9.16) and forming (according to (6.14))

$$l_1(p) = \xi_1(p + 1, 0) \mp i\xi_2(p + 1, 0)$$

and

$$l_2(p) = \pm\frac{i}{2}\xi_2(p+1,0) + \frac{\partial}{\partial s}\xi_1(p+1,0) \mp i\frac{\partial}{\partial s}\xi_2(p+1,0),$$

we obtain a solution for (9.2) *in Floquet form for the case* $v^2 = -4$ *as*

$$\sum_{-\infty}^{+\infty} [l_1(p), l_2(p)]z^p \begin{bmatrix} 1 & \log z/2\pi i \\ 0 & 1 \end{bmatrix}.$$

While we have constructed the Laurent coefficients L_p by the preceding infinite series for each integer p, it is interesting and useful to observe that because of the symmetries and special form of (9.2), it is only required to calculate a few such (transcendental) quantities in this manner; the remaining ones can then all be calculated recursively from the difference equation (0.6) and the symmetries as we now describe in the "main" case $v^2 \neq 0, -4$ (which corresponds to the nonperiodic case for (9.3)).

Let $X(z) = L(z)z^M$ denote the solution in Floquet form constructed above, where for definiteness we take

$$M = \operatorname{diag}\{\mu_1, -\mu_1\},$$

where $\mu_1 = \mu + k$, with μ denoting any *fixed* solution of $2i\cos\pi\mu = v$ and k denoting an integer. Then $2i\cos\pi\mu_1 = (-1)^k v$ and we will see below how the integer k affects the parity of the functions in $L(z)$. Observe that in our case ($v^2 \neq 0, -4$), $2\mu_1$ is not an integer.

Since

$$\operatorname{diag}\{1,-1\}X(z\,e^{\pi i}) = \operatorname{diag}\{1,-1\}L(-z)\,e^{\pi i M}z^M$$

is also a fundamental solution matrix for (9.2), there exists a constant, invertible matrix R such that

$$\operatorname{diag}\{1,-1\}L(-z)\,e^{\pi i M}z^M = L(z)z^M R,$$

from which it follows that $z^M R z^{-M}$ is single-valued. Since $2\mu_1$ is not an integer, it follows (see [10, p. 38]) that R is diagonal and

(9.19) $$\operatorname{diag}\{1,-1\}L(-z) = L(z)D \quad \text{where } D = \operatorname{diag}\{d_1, d_2\} = R\,e^{-\pi i M}.$$

Letting $L(z) = (\sum_{-\infty}^{+\infty} l_{ij}(p)z^p)$, $1 \leqq i, j \leqq 2$, it follows that for each integer p,

(9.20) $$(-1)^p l_{1j}(p) = d_j l_{1j}(p) \quad \text{and} \quad (-1)^{p+1} l_{2j}(p) = d_j l_{2j}(p),$$

$j = 1, 2$. Moreover, from Jacobi's identity one has $\det X(z) = \det L(z) = cz^{-1}$ for some nonzero constant c; hence from (9.19) $\det D = d_1 d_2 = 1$. Since neither column of $L(z)$ can be identically zero, it follows that either $d_1 = d_2 = 1$ and

(9.21) $$l_{1j}(2p+1) = 0, \quad l_{2j}(2p) = 0 \quad \text{for } j = 1, 2 \text{ and all } p$$

or $d_1 = d_2 = -1$ and

(9.22) $$l_{1j}(2p) = 0, \quad l_{2j}(2p+1) = 0 \quad \text{for } j = 1, 2 \text{ and all } p.$$

Which one of these cases occurs depends upon $k = \mu_1 - \mu$ being even or odd, as we will now see. From (9.17), setting $t = 0$ and recalling that in this event one must have $\arg(-ih) = \arg(ih) + \pi$ according to Theorem 2, we find

$$\xi_2(s,0) = \operatorname{diag}\{1,-1\}\,e^{i\pi(s-1/2)}\xi_1(s,0).$$

Using (9.18) and (9.15), it follows that

(9.23) $$l_1(p) = e^{i\pi\mu}\operatorname{diag}\{1 + e^{i\pi(p+\mu_1-\mu)}, 1 - e^{i\pi(p+\mu_1-\mu)}\}\xi_1(p+\mu_1+1,0);$$

hence (9.21) holds iff $\mu_1 - \mu$ is even while (9.22) holds iff $\mu_1 - \mu$ is odd.

The difference equation (0.6) for the first column of L_p in this case reduces to

$$[(p+\mu)^2-\tilde{\lambda}]l_{11}(p)+h^2l_{11}(p-2)+h^2l_{11}(p+2)=0 \tag{9.24}$$

and

$$l_{21}(p)=(\mu+p+1)l_{11}(p+1). \tag{9.25}$$

Hence one sees that if $l_{11}(p)$ and $l_{11}(p+2)$ are known for some value of p such that at least one of them is different from zero, then all the coefficients in the first column of $X(z)$ can be calculated recursively.

To generate the second column, one can use the fact that

$$\operatorname{diag}\{1, -1/z^2\}X(1/z)$$

is also a fundamental solution matrix, hence there exists an invertible, constant matrix P such that

$$\operatorname{diag}\{1, -1/z^2\}L(1/z)z^{-M}=L(z)z^MP.$$

From the single-valuedness of z^MPz^M and recalling 2μ is not an integer, one sees $\operatorname{diag} P=0$ and

$$\operatorname{diag}\{1, -1/z^2\}L(1/z)=L(z)P \quad \text{with } P=\begin{bmatrix}0 & \rho_2\\ \rho_1 & 0\end{bmatrix}. \tag{9.26}$$

Setting $z=1$ and taking the determinant, we have $\rho_1\rho_2=1$. Hence from (9.25) it follows that

$$l_{11}(p)=\rho_1l_{12}(-p) \quad \text{and} \quad -l_{21}(p)=\rho_1l_{22}(-p-2) \tag{9.27}$$

for all integers p.

Since for all p sufficiently large, it follows from the asymptotic (3.3) that the first component of $\xi_1(p+\mu_1+1, 0)$ is different from zero, then from the preceding discussion of the two cases (9.21) and (9.22), if we would select $k=0$, then $l_{11}(2n)\neq 0$ for all n sufficiently large. From (9.24) one sees that if two values $l_{11}(p)$, $l_{11}(p+2)$ would be zero, all the $l_{11}(p+2m)$ would be zero, hence if we are in case (9.21) then at least one of $l_{11}(p)$, $l_{11}(p+2)$ is different from zero for every even integer p. If $l_{11}(p)\neq 0$, then ρ_1 can be determined from (9.27) (by calculating $l_{12}(-p)$) and then all the other entries of the second column of $L(z)$ as well. On the other hand, if we just want *some* fundamental solution matrix in Floquet form, we could select $\rho_1=\rho_2=1$, which corresponds to multiplying the second column of $L(z)$ by a scalar, nonzero, constant (and also corresponds to changing the second column of the selected Ω in (9.15) by the same scalar factor). Since that Ω was not selected for any special purpose other than concreteness, if one just wants *a* fundamental solution in Floquet form, that point of view would be appropriate.

Hence we see that to calculate some solution for (9.2) in Floquet form according to this method, one must calculate three scalar quantities v, $l_{11}(p)$, $l_{11}(p+2)$, which are defined as the limits of infinite series. All other quantities in a Floquet solution can be recursively generated from these.

REFERENCES

[1] W. BALSER, W. B. JURKAT AND D. A. LUTZ, *On the reduction of connection problems for differential equations with an irregular singular point to ones with only regular singularities*, I, this Journal, 12 (1981), pp. 691–721.

[2] W. BALSER, W. B. JURKAT AND D. A. LUTZ, *Transfer of connection problems for first level solutions of meromorphic differential equations, and associated Laplace transforms*, J. Reine Angew. Math., 344 (1983), pp. 149-170.

[3] ———, *A general theory of invariants for meromorphic differential equations*, I, *formal invariants*, Funk. Ekvac., 22 (1979), pp. 197-221; II, *proper invariants*, Funk. Ekvac., 22 (1979), pp. 257-283.

[4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[5] G. DOETSCH, *Handbuch der Laplace-Transformation*, Vols. I, II, Birkhäuser-Verlag, Basel, 1972.

[6] F. R. GANTMACHER, *Theory of Matrices*, Vol. I, Chelsea, New York, 1959.

[7] F. L. HINTON, *Stokes multipliers for a class of ordinary differential equations*, J. Math. Phys., 20 (1979), pp. 2036-2046.

[8] L. HOPF, *Fortsetzungsrelationen bei den Lösungen gewöhnlicher linearer Differentialgleichungen*, Math. Ann., 111 (1935), pp. 678-712.

[9] M. HUKUHARA, *Développments asymptotiques des solutions principales d'un systeme différential linéaire du type hypergéométrique*, Tokyo J. Math., 5 (1982), pp. 491-499.

[10] W. B. JURKAT, *Meromorphe Differentialgleichungen*, Lecture Notes in Mathematics 637 (entire volume), Springer-Verlag, Berlin-Heidelberg-New York, 1978.

[11] W. B. JURKAT, D. A. LUTZ AND A. PEYERIMHOFF, *Birkhoff invariants and effective calculations for meromorphic linear differential equations*, I, J. Math. Anal. Appl., 53 (1976), pp. 438-470.

[12] H. W. KNOBLOCH, *Zusammenhänge zwischen konvergenten und asymptotischen Entwicklungen bei Lösungen linearer Differentialsysteme vom Range Eins*, Math. Anna., 134 (1958), pp. 260-288.

[13] H. VON KOCH, *Sur une application des déterminants infinis á la theorie des équations différentielles linéaires*, Acta. Math., 15 (1891/92).

[14] M. KOHNO, *A two point connection problem*, Hiroshima Math. J., 9 (1979), pp. 61-135.

[15] K. OKUBO, *A global representation of a fundamental set of solutions and a Stokes phenomenon for a system of ordinary linear differential equations*, J. Math. Soc. Japan, 15 (1963), pp. 268-288.

[16] N. E. NÖRLUND, *Vorlesungen über Differenzenrechnung*, Chelsea, New York, 1954.

[17] R. SCHÄFKE, *Über das globale analytische Verhalten der Lösungen der über die Laplacetransformation zusammenhängenden Differentialgleichungen* $tx' = (A + tB)x$ *und* $(s - B)v' = (\rho - A)v$, Doctoral Dissertation, Universität Essen, West Germany, April 1979.

[18] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

[19] A. ERDÉLYI, *Über die Integration der Mathieuschen Differentialgleichung durch Laplacesche Integrale*, Math. Z., 41 (1936), pp. 653-664.

[20] ———, *Bemerkungen zur Integration der Mathieuschen Differentialgleichung durch Laplacesche Integrale*, Compositio Math., 5 (1938), pp. 435-446.

[21] W. B. JURKAT, D. A. LUTZ AND A. PEYERIMHOFF, *Invariants and canonical forms for meromorphic second order differential equations*, in New Developments in Differential Equations, W. Eckhaus, ed., North-Holland, Amsterdam, 1976, pp. 181-187.

[22] J. MEIXNER AND F. W. SCHÄFKE, *Mathieusche Funktionen und Sphäroidfunktionen*, Springer-Verlag, Grundlehren Math. Wiss. Band LXXI, 1954.

[23] R. SCHÄFKE, *Über das globale Verhalten der Normallösungen von* $\chi'(t) = (B + t^{-1}A)\chi(t)$ *und zweier Arten von assoziierten Funktionen*, Math. Nachr., 121 (1985), pp. 123-145.

[24] Y. SIBUYA, private communication.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

013

ASYMPTOTIC INTEGRATION OF A PERTURBED CONSTANT COEFFICIENT DIFFERENTIAL EQUATION UNDER MILD INTEGRAL SMALLNESS CONDITIONS*

WILLIAM F. TRENCH†

Abstract. The problem of asymptotic behavior of solutions of an nth order linear differential equation is reconsidered, and a result obtained by Hartman and Wintner under integral smallness conditions on the perturbing terms which require absolute integrability is shown to hold under weaker integrability conditions requiring only ordinary (perhaps conditional) convergence of some of the improper integrals that occur. The estimates of the asymptotic behavior of solutions of the perturbed equation are also sharper than in the classical result.

Key words. linear perturbations, asymptotic behavior, integral smallness conditions

AMS(MOS) subject classifications. 34C10, 34D10, 34E10

1. Introduction.

We consider the scalar equation

$$x^{(n)}+a_1x^{(n-1)}+\cdots+a_kx^{(n-k)}+P_1(t)x^{(n-1)}+\cdots+P_n(t)x=f(t), \qquad t>0, \tag{1.1}$$

where $a_1,\cdots,a_k$ are complex constants, with

$$1\leqq k\leqq n-1, \qquad a_k\neq 0. \tag{1.2}$$

It is assumed through that $P_1,\cdots,P_n$, and f are complex-valued and continuous on $(0,\infty)$. We give conditions on them which imply that (1.1) has a solution which behaves asymptotically like a given polynomial of degree $<n-k$.

The following theorem of Hartman and Wintner [1, Thm. 17.3, p. 316] addresses this question. (We use "o" and "O" in the standard way to denote behavior as $t\to\infty$.)

THEOREM 1. *Suppose that the polynomial*

$$Q(\lambda)=\lambda^k+a_1\lambda^{k-1}+\cdots+a_k$$

has no purely imaginary zeros, and that

$$\int^{\infty}|P_j(t)|t^q\,dt<\infty, \qquad 1\leqq j\leqq k+1, \tag{1.3}$$

and

$$\int^{\infty}|P_j(t)|t^{j-k-1+q}\,dt<\infty, \qquad k+2\leqq j\leqq n, \tag{1.4}$$

for some $q\geqq 0$. Then, for each $l=0,1,\cdots,n-k-1$, the equation

$$x^{(n)}+a_1x^{(n-1)}+\cdots+a_kx^{(n-k)}+P_1(t)x^{(n-1)}+\cdots+P_n(t)x=0, \qquad t>0, \tag{1.5}$$

has a solution x_l such that

$$\left(x_l(t)-\frac{t^l}{l!}\right)^{(r)}=\begin{cases} o(t^{l-r-q}), & 0\leqq r\leqq n-k-1,\\ o(t^{-n+l+k+1-q}), & n-k\leqq r\leqq n-1.\end{cases}$$

Prevatt [3] has obtained the conclusion of Theorem 1 under weaker integrability conditions on $|P_1|,\cdots,|P_n|$, and Hartman [2] has recently extended Prevatt's results

* Received by the editors March 31, 1986; accepted for publication (in revised form) December 22, 1986.
† Department of Mathematics, Trinity University, San Antonio, Texas 78284.

to the case where $Q(\lambda)$ may have purely imaginary zeros. Here we retain the assumption that $Q(\lambda)$ has no purely imaginary zeros, and we obtain results which imply the conclusion of Theorem 1 under integral smallness conditions on $P_1, \cdots, P_n$ and f which allow conditional convergence of some (in some cases all) of the improper integrals involved. We also give more precise estimates of the asymptotic behavior of the desired solutions.

The results obtained here are anaolgous to those obtained in [8] for the equation

$$x^{(n)} + P_1(t)x^{(n-1)} + \cdots + P_n(t)x = f(t) \tag{1.6}$$

(see also [7]); however, the condition (1.2) necessitates a distinctly nontrivial extension of the methods used in [7] and [8]. (For example, see Lemma 1 and its proof.)

In work related to the present paper in the sense that the integrability conditions on $P_1, \cdots, P_n$, and f are stated in terms of possibly conditional convergence, Šimša [4]-[6], has studied (1.1) with $k = n$, regarding it as a perturbation of the constant coefficient equation

$$x^{(n)} + a_1 x^{(n-1)} + \cdots + a_n x = 0.$$

The author [9] has also considered this problem.

2. The main theorem. It is to be understood below that improper integrals appearing in hypotheses are assumed to converge, and that the convergence may be conditional unless, of course, the integrand is necessarily nonnegative.

It is convenient to collect some technical definitions in the following standing assumption, which holds throughout the paper.

Assumption A. Let

$$Q(\lambda) = (\lambda - \lambda_1)^{d_1} \cdots (\lambda - \lambda_L)^{d_L} \tag{2.1}$$

where $\lambda_l = \mu_l + i\nu_l$ are distinct, $\mu_l \neq 0$ $(1 \leqq l \leqq L)$, and $\mu_1 \leqq \mu_2 \leqq \cdots \leqq \mu_L$. Let N be the unique integer in $\{1, \cdots, L+1\}$ such that

$$\mu_l < 0 \quad \text{if } 1 \leqq l \leqq N-1 \tag{2.2}$$

and $\mu_l > 0$ if $N \leqq l \leqq L$. Suppose that p is a given polynomial of degree $< n-k$, and define

$$g = f - \sum_{j=k+1}^{n} P_j p^{(n-j)}. \tag{2.3}$$

Let m be an integer in $\{0, \cdots, n-k-1\}$. Let ϕ be continuous, positive, and nonincreasing on $[a, \infty)$ for some $a > 0$. If $m \neq 0$, suppose that $t^{\gamma}\phi(t)$ is nondecreasing for some $\gamma < 1$. If $N \geqq 2$ (so that (2.2) is nonvacuous), let there be a number α such that $0 < \alpha < -\mu_{N-1}$ and $e^{\alpha t}t^{-n+m+k+1}\phi(t)$ is nondecreasing on $[a, \infty)$.

The following is our main theorem.

THEOREM 2. *Suppose that Assumption* A *holds, that*

$$\int_t^{\infty} s^{n-m-k-1} g(s)\, ds = O(\phi(t)), \tag{2.4}$$

and that

$$I_j(t) = \int_t^{\infty} |P_j(s)|\phi(s)\, ds = o(\phi(t)), \qquad 1 \leqq j \leqq k+1. \tag{2.5}$$

Suppose further that

$$F_j(t)=\int_t^\infty P_j(s)\,ds=o(t^{-j+k+1}),\qquad k+2\leqq j\leqq n, \tag{2.6}$$

and that

$$I_j(t)=\int_t^\infty |F_j(s)|s^{j-k-2}\phi(s)\,ds=o(\phi(t)),\qquad k+2\leqq j\leqq n. \tag{2.7}$$

Then (1.1) *has a solution* $\hat{x}$ *such that*

$$\hat{x}^{(r)}(t)-p^{(r)}(t)=\begin{cases}O(\phi(t)t^{m-r}), & 0\leqq r\leqq n-k-1,\\ O(\phi(t)t^{-n+m+k+1}), & n-k\leqq r\leqq n-1.\end{cases} \tag{2.8}$$

Moreover, if (2.4) *can be replaced by*

$$\int_t^\infty s^{n-m-k-1}g(s)\,ds=o(\phi(t)), \tag{2.9}$$

then (2.8) *can be replaced by*

$$\hat{x}^{(r)}(t)-p^{(r)}(t)=\begin{cases}o(\phi(t)t^{m-r}), & 0\leqq r\leqq n-k-1,\\ o(\phi(t)t^{-n+m+k+1}), & n-k\leqq r\leqq n-1.\end{cases} \tag{2.10}$$

The proof uses the Banach contraction principle. It is convenient to introduce the new dependent variable $h=x-p$, which transforms (1.1) into

$$Q(D)h^{(n-k)}=g-Mh, \tag{2.11}$$

with g as in (2.3) and

$$Mh=\sum_{j=1}^{n}P_jh^{(n-j)}. \tag{2.12}$$

We will construct a transformation $\mathcal{T}$ which, for t_0 sufficiently large, is a contraction of the Banach space $B(t_0)$ consisting of functions h in $C^{(n-1)}[t_0,\infty)$ such that

$$h^{(r)}(t)=\begin{cases}O(\phi(t)t^{m-r}), & 0\leqq r\leqq n-k-1,\\ O(\phi(t)t^{-n+m+k+1}), & n-k\leqq r\leqq n-1,\end{cases} \tag{2.13}$$

with norm

$$\|h\|=\sup_{t\geqq t_0}(\phi(t))^{-1}\left[\sum_{r=0}^{n-k-1}t^{r-m}|h^{(r)}(t)|+t^{n-m-k-1}\sum_{r=n-k}^{n-1}|h^{(r)}(t)|\right]. \tag{2.14}$$

For reference below, we define

$$B_0(t_0)=\{h\in B(t_0)|\ (2.13)\text{ holds with ``}O\text{'' replaced by ``}o\text{''}\}. \tag{2.15}$$

If $\hat{h}$ is a solution of (2.11) which is in $B(t_0)$, then the function

$$\hat{x}=p+\hat{h} \tag{2.16}$$

is a solution of (1.1) on $[t_0,\infty)$ (which can be continued over $(0,\infty)$), and $\hat{x}$ has the desired asymptotic behavior (2.8); moreover, if $\hat{h}\in B_0(t_0)$, then $\hat{x}$ satisfies (2.10). Therefore, we wish to construct $\mathcal{T}$ so that if $\mathcal{T}\hat{h}=\hat{h}$, then $\hat{h}$ satisfies (2.11).

To this end, let $A_1, \cdots, A_L$ be the uniquely defined polynomials such that $\deg A_l < d_l$ (see (2.1)) and

$$\sum_{l=1}^{L} [A_l(t)e^{\lambda_l t}]^{(r)}|_{t=0} = \delta_{r,k-1}, \qquad 0 \leqq r \leqq k-1. \tag{2.17}$$

If $v \in C[t_0, \infty)$, define $\mathscr{L}_1 v$ formally by

$$(\mathscr{L}_1 v)(t) = \sum_{l=1}^{N-1} \int_{t_0}^{t} A_l(t-\tau)e^{\lambda_l(t-\tau)}v(\tau)\,d\tau - \sum_{l=N}^{L} \int_{t}^{\infty} A_l(t-\tau)\,e^{\lambda_l(t-\tau)}v(\tau)\,d\tau.$$

Then formal differentiation and (2.17) imply that

$$\begin{aligned}(\mathscr{L}_1 v)^{(r)}(t) = &\sum_{l=1}^{N-1} \int_{t_0}^{t} A_{lr}(t-\tau)\,e^{\lambda_l(t-\tau)}v(\tau)\,d\tau \\ &- \sum_{l=N}^{L} \int_{t}^{\infty} A_{lr}(t-\tau)\,e^{\lambda_l(t-\tau)}v(\tau)\,d\tau, \qquad 0 \leqq r \leqq k-1,\end{aligned} \tag{2.18}$$

and that

$$Q(D)\mathscr{L}_1 v = v, \tag{2.19}$$

where A_{lr} is the polynomial defined by

$$A_{lr}(t) = e^{-\lambda_l t}[A_l(t)\,e^{\lambda_l t}]^{(r)}, \qquad 0 \leqq r \leqq k-1.$$

If $w \in C[t_0, \infty)$, define $\mathscr{L}_2 w$ formally by

$$(\mathscr{L}_2 w)(t) = \int_{t}^{\infty} \frac{(t-s)^{n-k-1}}{(n-k-1)!}\,w(s)\,ds \quad \text{if } m = 0 \tag{2.20}$$

or by

$$(\mathscr{L}_2 w)(t) = \int_{t_0}^{t} \frac{(t-\lambda)^{m-1}}{(m-1)!}\,d\lambda \int_{\lambda}^{\infty} \frac{(\lambda-s)^{n-k-m-1}}{(n-k-m-1)!}\,w(s)\,ds \quad \text{if } 1 \leqq m \leqq n-k-1. \tag{2.21}$$

In either case,

$$(\mathscr{L}_2 w)^{(n-k)} = -w. \tag{2.22}$$

Now define

$$G = \mathscr{L}_2(\mathscr{L}_1 g) \quad \text{(see (2.3))}, \tag{2.23}$$

$$\mathscr{L}h = \mathscr{L}_2(\mathscr{L}_1(Mh)) \quad \text{(see (2.12))}, \tag{2.24}$$

and

$$\mathscr{T}h = -G + \mathscr{L}h. \tag{2.25}$$

Formal manipulations using (2.19) and (2.22) show that

$$Q(D)(\mathscr{T}h)^{(n-k)} = g - Mh;$$

therefore, a fixed point (function) $\hat{h}$ of $\mathscr{T}$ satisfies (2.11). We will show that $\mathscr{T}$ is a contraction of $B(t_0)$, and therefore has a fixed point in $B(t_0)$, provided that t_0 is sufficiently large. It is convenient to present the lengthy proof of this assertion in a series of lemmas.

LEMMA 1. *Suppose that v is complex-valued and continuous on $[t_0, \infty)$ with $t_0 \geqq a > 0$, and that $\int^\infty t^q v(t)\,dt$ converges for some nonnegative integer q. Let*

$$\psi(t) = \sup_{\tau \geqq t} \left| \int_\tau^\infty s^q v(s)\,ds \right|. \tag{2.26}$$

Let $\lambda = \mu + i\nu$ be a complex number and X be a polynomial. Then

(i) *If $\mu > 0$, the functions*

$$f_1(t) = \int_t^\infty X(t-\tau)\, e^{\lambda(t-\tau)} v(\tau)\,d\tau \tag{2.27}$$

and

$$f_2(t) = \int_t^\infty s^q\,ds \int_s^\infty X(s-\tau)\, e^{\lambda(s-\tau)} v(\tau)\,d\tau = \int_t^\infty s^q f_1(s)\,ds \tag{2.28}$$

are defined on $[t_0, \infty)$ and satisfy the inequalities

$$|f_1(t)| \leqq K_1 t^{-q} \psi(t), \qquad t \geqq t_0, \tag{2.29}$$

$$|f_2(t)| \leqq K_2 \psi(t), \qquad t \geqq t_0, \tag{2.30}$$

where K_1 and K_2 are constants which depend only on λ and X.

(ii) *If $\mu < 0$, suppose that $\psi(t) = 0(\phi(t))$, where ϕ is nonincreasing and continuous on $[a, \infty)$, and $e^{\alpha t} t^{-q} \phi(t)$ in nondecreasing on $[a, \infty)$ for some α such that*

$$0 < \alpha < -\mu. \tag{2.31}$$

Let

$$\psi_1(t) = \sup_{\tau \geqq t} \frac{\psi(t)}{\phi(t)}, \tag{2.32}$$

and define

$$f_3(t) = \int_{t_0}^t X(t-\tau)\, e^{\lambda(t-\tau)} v(\tau)\,d\tau. \tag{2.33}$$

Then

$$|f_3(t)| \leqq K_3 \psi_1(t_0) t^{-q} \phi(t), \qquad t \geqq t_0, \tag{2.34}$$

where K_3 is a constant depending only on X and λ, and the function

$$f_4(t) = \int_t^\infty s^q\,ds \int_{t_0}^s X(s-\tau)\, e^{\lambda(s-\tau)} v(\tau)\,d\tau = \int_t^\infty s^q f_3(s)\,ds \tag{2.35}$$

is defined on $[t_0, \infty)$, and it satisfies the inequality

$$|f_4(t)| \leqq K_4 \psi_1(t_0) \phi(t), \qquad t \geqq t_0, \tag{2.36}$$

where K_4 is a constant depending only on X and λ.

Finally, if

$$\psi(t) = o(\phi(t)), \tag{2.37}$$

then

$$f_3(t) = o(t^{-q}\phi(t)) \tag{2.38}$$

and

$$f_4(t) = o(\phi(t)). \tag{2.39}$$

The lengthy proof of this lemma is given in § 4.

(Note that because of (2.29) and (2.30), (2.37) implies that $f_1(t) = o(t^{-q}\phi(t))$ and $f_2(t) = o(\phi(t))$.)

LEMMA 2. *Suppose that ϕ and m are as in Assumption* A *and $w \in C[t_0, \infty)$ for some $t_0 \geqq a$. Suppose also that $\int^\infty t^{n-m-k-1}w(t)\,dt$ converges, and that*

$$\int_t^\infty s^{n-m-k-1}w(s)\,ds = O(\phi(t)). \tag{2.40}$$

Define

$$\rho(t) = \sup_{\tau \geqq t} (\phi(\tau))^{-1} \left| \int_\tau^\infty s^{n-k-m-1}w(s)\,ds \right|. \tag{2.41}$$

Then $\mathscr{L}_2 w \in C^{(n-k)}[t_0, \infty)$ (see (2.20) *and* (2.21)), *and there is a constant K which does not depend on w or t_0 such that*

$$|(\mathscr{L}_2 w)^{(r)}(t)| \leqq K\rho(t_0)\phi(t)t^{m-r}, \qquad t \geqq t_0, \quad 0 \leqq r \leqq n-k-1. \tag{2.42}$$

Moreover, if

$$\lim_{t\to\infty} \rho(t) = 0, \tag{2.43}$$

then

$$(\mathscr{L}_2 w)^{(r)}(t) = o(\phi(t)t^{m-r}), \qquad 0 \leqq r \leqq n-k-1. \tag{2.44}$$

This lemma follows immediately from Lemma 1 of [8]. Since its conclusion follows trivially from the assumption that

$$\int^\infty s^{n-k-m-1}|w(s)|\,ds < \infty,$$

it is important to emphasize here that the covergence in (2.40) may be conditional. (Note: The existence of the constant γ in Assumption A is required for this lemma.)

LEMMA 3. *Suppose that v is complex-valued and continuous on $[t_0, \infty)$ with $t_0 \geqq a$, and that $\int^\infty t^{n-m-k-1}v(t)\,dt$ converges. Let*

$$\psi(t) = \sup_{\tau \geqq t} \left| \int_\tau^\infty s^{n-m-k-1}v(s)\,ds \right| = O(\phi(t)), \tag{2.45}$$

and define ψ_1 as in (2.32). *Then the function*

$$u = \mathscr{L}_2(\mathscr{L}_1 v) \tag{2.46}$$

is in $B(t_0)$, and

$$\|u\| \leqq W\psi_1(t_0), \tag{2.47}$$

where W is a constant independent of t_0 and v. Moreover, if

$$\psi(t) = o(\phi(t)), \tag{2.48}$$

then

$$u \in B_0(t_0) \quad (\textit{see } (2.15)). \tag{2.49}$$

Proof. From (2.18), $\mathscr{L}_1 v$ and its first $k-1$ derivatives are linear combinations of integrals of the forms (2.27) and (2.33) with $X = A_{lr}$: hence, Lemma 1 with $q = n-m-k-1$ (specifically, (2.29), (2.32) and (2.34)) implies that $\mathscr{L}_1 v \in C^{(k-1)}[t_0, \infty)$, and that

$$|(\mathscr{L}_1 v)^{(j)}(t)| \leqq \alpha_1 \psi_1(t_0)\phi(t)t^{-n+m+k+1}, \qquad 0 \leqq j \leqq k-1, \tag{2.50}$$

where α_1 is a constant independent of t_0 and v. Lemma 1 also implies that $\int^\infty t^{n-m-k-1}(\mathscr{L}_1 v)(t)\,dt$ converges (since it is a linear combination of integrals of the forms (2.28) and (2.35), again with $q = n-m-k-1$ and $X = A_{lr}$), and that

$$\left| \int_t^\infty s^{n-m-k-1}(\mathscr{L}_1 v)(s)\,ds \right| \leqq \alpha_2 \psi_1(t_0)\phi(t), \qquad t \geqq t_0, \tag{2.51}$$

where α_2 is a constant independent of t_0 and v. (See (2.30), (2.32), and (2.36).)

Now we apply Lemma 2 with $w = \mathscr{L}_1 v$. Then (2.51) implies (2.40) and (2.41), with $\rho(t) \leqq \alpha_2 \psi_1(t_0)$. Recalling (2.46), we now see from (2.42) with $w = \mathscr{L}_1 v$ that

$$|u^{(r)}(t)| \leqq K\alpha_2 \psi_1(t_0)\phi(t)t^{m-r}, \qquad 0 \leqq r \leqq n-k-1. \tag{2.52}$$

Moreover, since

$$u^{(n-k+j)} = -(\mathscr{L}_1 v)^{(j)}, \qquad 0 \leqq j \leqq k-1 \tag{2.53}$$

(from (2.22) with $w = \mathscr{L}_1 v$), (2.50) implies that

$$|u^{(r)}(t)| \leqq \alpha_1 \psi_1(t_0)\phi(t)t^{-n+m+k+1}, \qquad n-k \leqq r \leqq n-1. \tag{2.54}$$

Now (2.14), (2.52) and (2.54) imply (2.47), with $W = \max\{\alpha_1, K\alpha_2\}$.

It remains only to show that (2.48) implies (2.49). From the closing paragraph of Lemma 1, (2.48) implies that

$$(\mathscr{L}_1 v)^{(j)}(t) = o(\phi(t)t^{-n+m+k+1}), \qquad 0 \leqq j \leqq k-1,$$

and therefore

$$u^{(r)}(t) = o(\phi(t)t^{-n+m+k+1}), \qquad n-k \leqq r \leqq n-1, \tag{2.55}$$

because of (2.53). The closing paragraph of Lemma 1 also implies that

$$\int_t^\infty s^{n-m-k-1}(\mathscr{L}_1 v)(s)\,ds = o(\phi(t)),$$

because of (2.48). Therefore, (2.43) holds if $w = \mathscr{L}_1 v$ in (2.41), and so

$$u^{(r)}(t) = o(\phi(t)t^{m-r}), \qquad 0 \leqq r \leqq n-k-1, \tag{2.56}$$

from (2.44). Since (2.55) and (2.56) are equivalent to (2.49), this completes the proof of Lemma 3.

LEMMA 4. *Suppose that the assumptions of Theorem 2 hold, and let $\mathscr{L}$ be as defined in* (2.24). *Suppose also that $h \in B(t_0)$ for some $t_0 \geqq a$. Then $\mathscr{L}h \in B_0(t_0)$ and*

$$\|\mathscr{L}h\| \leqq W\sigma(t_0)\|h\|, \tag{2.57}$$

where σ is defined on $[a, \infty)$,

$$\lim_{t\to\infty} \sigma(t) = 0, \tag{2.58}$$

and W is as in (2.47).

Proof. We first consider the integral

$$J(t;h)=\int_t^\infty s^{n-m-k-1}(Mh)(s)\,ds$$

$$=\sum_{j=1}^{n}\int_t^\infty s^{n-m-k-1}P_j(s)h^{(n-j)}(s)\,ds. \tag{2.59}$$

We will show that the integrals in this sum converge, and estimate them. We remind the reader that $\|h\|$ is defined in (2.14).

From (2.5),

$$\left|\int_t^\infty s^{n-m-k-1}P_j(s)h^{(n-j)}(s)\,ds\right|\leqq\|h\|I_j(t),\qquad 1\leqq j\leqq k+1. \tag{2.60}$$

If $k+2\leqq j\leqq n$, then integration by parts yields

$$\int_t^\infty s^{n-m-k-1}P_j(s)h^{(n-j)}(s)\,ds$$

$$=t^{n-m-k-1}h^{(n-j)}(t)F_j(t)+\int_t^\infty F_j(s)[s^{n-m-k-1}h^{(n-j)}(s)]'\,ds\quad\text{(see (2.6))}. \tag{2.61}$$

To justify this we first observe that

$$\lim_{T\to\infty}T^{n-m-k-1}h^{(n-j)}(T)F_j(T)=0,\qquad k+2\leqq j\leqq n,$$

because of (2.6) and (2.13). Moreover, since (2.13) and (2.14) imply that

$$|[s^{n-m-k-1}h^{(n-j)}(s)]'|\leqq(n-m-k)\|h\|\phi(s)s^{j-k-2},\qquad k+2\leqq j\leqq n, \tag{2.62}$$

(2.7) implies that the integral on the right side of (2.61) converges (absolutely). We now conclude that $J(t;h)$ exists on $[t_0,\infty)$ if $h\in B(t_0)$; moreover, (2.59), (2.60), (2.61) and (2.62) imply that $|J(t;h)|\leqq\|h\|\Gamma(t)$, where

$$\Gamma(t)=\sum_{j=1}^{k+1}I_j(t)+(n-m-k)\sum_{j=k+1}^{n}I_j(t)+\phi(t)\sum_{j=k+2}^{n}t^{j-k-1}|F_j(t)|.$$

(See (2.6) and (2.7).) Now define $\sigma(t)=\sup_{\tau\geqq t}\Gamma(\tau)/\phi(\tau)$, and note that σ satisfies (2.58), because of (2.5), (2.6) and (2.7).

We now apply Lemma 3 with $v=Mh$; then the function ψ defined in (2.45) satisfies the inequality

$$\psi(t)\leqq\|h\|\Gamma(t)=o(\phi(t)).$$

Since $u=\mathscr{L}h$ in (2.46) when $v=Mh$ (see (2.24)), we conclude that $\mathscr{L}h\in B_0(t_0)$, and (2.47) with $\psi_1=\sigma\|h\|$ implies (2.57). This proves Lemma 4.

We can now complete the proof of Theorem 2. From (2.4) and Lemma 3 (with $v=g$), G as defined in (2.23) is also in $B(t_0)$. Now (2.25) and Lemma 4 imply that $\mathscr{T}(B(t_0))\subset B(t_0)$; moreover, $\mathscr{T}$ is obviously a contraction if $\mathscr{L}$ is, and the latter is so if

$$\sigma(t_0)<1/W\quad\text{(see (2.57))}. \tag{2.63}$$

From (2.58), we can choose t_0 to satisfy (2.63). By the contraction mapping principle, there is an $\hat{h}$ in $B(t_0)$ such that $\mathscr{T}\hat{h}=\hat{h}$; hence, $\hat{x}$ as defined in (2.16) satisfies (1.1) and (2.8).

Now suppose that (2.9) holds. Then Lemma 3 implies that $G \in B_0(t_0)$. Since $\mathscr{L}\hat{h} \in B_0(t_0)$ (by Lemma 4) and

$$\hat{h} = -G + \mathscr{L}\hat{h} \quad \text{(see (2.25))},$$

it now follows that $\hat{h} \in B_0(t_0)$. This and (2.16) imply (2.10), which completes the proof of Theorem 2.

3. Corollaries. If $P \in C[a, \infty)$ and $\int^\infty |P(t)|\, dt < \infty$, then obviously

$$\int_t^\infty |P(s)|\phi(s)\, ds = o(\phi(t))$$

if ϕ is nonincreasing. Also, if $\int^\infty t^\alpha |P(t)|\, dt < \infty$ for some $\alpha > 0$, then

$$\int_t^\infty P(s)\, ds = o(t^{-\alpha}) \quad \text{and} \quad \int^\infty t^{\alpha-1} \left| \int_t^\infty P(\tau)\, d\tau \right| dt < \infty$$

(see Corollary 3 of [8]), which in turn implies that

$$\int_t^\infty s^{\alpha-1} \left| \int_s^\infty P(\tau)\, d\tau \right| \phi(s)\, ds = o(\phi(t))$$

if ϕ is nonincreasing. Since the converses of these statements are false, the integrability conditions (1.3) and (1.4)—even with $q = 0$—are stronger than (2.5), (2.6) and (2.7).

We remind the reader that Assumption A is still in force.

COROLLARY 1. *Let l be an integer in $\{0, 1, \cdots, n-k-1\}$, and suppose that*

$$\int_t^\infty s^{j-k-1} P_j(s)\, ds = O(t^{-q}), \qquad n-l \leqq j \leqq n, \tag{3.1}$$

for some $q \geqq 0$ such that

$$q \neq 0, 1, \cdots, l. \tag{3.2}$$

Let

$$\int_t^\infty P_j(s)\, ds = o(t^{-j+k+1}), \qquad k+2 \leqq j \leqq n-l-1, \tag{3.3}$$

and define

$$\beta = \begin{cases} q - [q] & \text{if } l \geqq [q] \ (= \text{integer part of } q), \\ q - l & \text{if } l < [q]. \end{cases} \tag{3.4}$$

Finally, suppose that

$$\int_t^\infty |P_j(s)| s^{-\beta}\, ds = o(t^{-\beta}), \qquad 1 \leqq j \leqq k+1. \tag{3.5}$$

Then (1.5) *has a solution x_l such that*

$$\left(x_l(t) - \frac{t^l}{l!}\right)^{(r)} = \begin{cases} O(t^{l-r-q}), & 0 \leqq r \leqq n-k-1, \\ O(t^{l-n+k+1-q}), & n-k \leqq r \leqq n-1. \end{cases} \tag{3.6}$$

Moreover, if (3.1) *can be replaced by*

$$\int_t^\infty P_j(s)\, ds = o(t^{-j+k+1-q}), \qquad n-l \leqq j \leqq n, \tag{3.7}$$

then (3.6) *can be replaced by*

$$\left(x_l(t)-\frac{t^l}{l!}\right)^{(r)}=\begin{cases}o(t^{l-r-q}), & 0\leqq r\leqq n-k-1,\\ o(t^{l-n+k+1-q}), & n-k\leqq r\leqq n-1.\end{cases} \tag{3.8}$$

Proof. We start by observing that if $0<a<b$ and

$$\int_t^\infty P(s)\,ds=O(t^{-b}), \tag{3.9}$$

then

$$\int_t^\infty s^aP(s)\,ds=O(t^{a-b}). \tag{3.10}$$

If $f=0$ (as in (1.5)) and $p(t)=t^l/l!$, then

$$g(t)=-\sum_{j=n-l}^{n}P_j(t)\frac{t^{l-n+j}}{(l-n+j)!} \tag{3.11}$$

(see (2.3)); hence, (3.1) implies that

$$\int_t^\infty s^{n-m-k-1}g(s)\,ds=O(t^{l-m-q})$$

if $l-m<q$. We now apply Theorem 2 with

$$m=\max\{0,\,l-[q]\} \tag{3.12}$$

and

$$\phi(t)=O(t^{l-m-q})=O(t^{-\beta}), \tag{3.13}$$

with β as in (3.4). (Note that if $m>0$, then $\beta=q-[q]<1$, and ϕ as in (3.13) satisfies Assumption A with $\gamma=\beta$.)

Now (3.1), (3.2) and (3.3) imply (2.6), and (3.5) is the same as (2.5) with $\phi(t)=t^{-\beta}$. Since (3.2) and (3.4) imply that $\beta>0$, (2.6) automatically implies (2.7) with ϕ as in (3.13), without any absolute integrability assumptions on $P_{k+2},\cdots,P_n$. Now Theorem 2 implies that (1.5) has a solution x_l such that

$$\left(x_l(t)-\frac{t^l}{l!}\right)^{(r)}=\begin{cases}O(t^{m-r-\beta}), & 0\leqq r\leqq n-k-1,\\ O(t^{m-n+k+1-\beta}), & n-k\leqq r\leqq n-1,\end{cases}$$

which, in view of (3.4) and (3.12), is equivalent to (3.6).

If (3.9) holds with "O" replaced by "o", then so does (3.10). From Theorem 2, this also implies the assertion concerning (3.7) and (3.8).

We now consider the exceptional cases where (3.2) is not satisfied.

COROLLARY 2. *Let q and l be integers, with $0\leqq q\leqq l\leqq n-k-1$. Suppose that the integrals*

$$\int^\infty t^{j-k-1+q}P_j(t)\,dt, \qquad n-l\leqq j\leqq n, \tag{3.14}$$

converge, that (3.3) *holds, that*

$$\int^\infty |P_j(t)|\,dt<\infty, \qquad 1\leqq j\leqq k+1, \tag{3.15}$$

and that

$$\int^\infty t^{j-k-2}|F_j(t)|\,dt<\infty \tag{3.16}$$

for $k+2\leqq j\leqq n-l-1$. *Then* (1.5) *has a solution* x_l *which satisfies* (3.8) *if* $q>0$. *If* $q=0$, *the same conclusion is valid if the above assumptions hold and, in addition,* (3.16) *also holds for* $n-l\leqq j\leqq n$.

Proof. Now (3.14) implies that $\int^\infty t^{n-m-k-1}g(t)\,dt$ converges (see (3.11)) with $m=l-q$, and that (3.7) holds. We apply Theorem 2 with $\phi=1$, in which case (2.5) is equivalent to (3.15). Also, (2.7) is then equivalent to (3.16) for $k+2\leqq j\leqq n$. Since the convergence of (3.14) implies (3.7), which automatically implies (3.16) for $n-l\leqq j\leqq n$, the proof is complete.

Corollaries 1 and 2 imply the following corollary, which in turn implies Theorem 1.

COROLLARY 3. *Suppose that the integrals*

$$\int^\infty t^{j-k-1+q}P_j(t)\,dt, \qquad k+1\leqq j\leqq n, \tag{3.17}$$

converge for some $q\geqq 0$. *Suppose also that*

$$\int_t^\infty |P_j(s)|s^{[q]-q}\,ds=o(t^{[q]-q}), \qquad 1\leqq j\leqq k+1. \tag{3.18}$$

Then the conclusions of Theorem 1 *hold if either* (i) $q>0$; *or* (ii) $q=0$ *and*

$$\int^\infty s^{j-k-2}\left|\int_s^\infty P_j(\tau)\,d\tau\right|ds<\infty, \qquad k+2\leqq j\leqq n. \tag{3.19}$$

It is important to note that (3.17) implies (3.19) if $q>0$; therefore, it is not necessary to assume (3.19) in this case.

Corollary 1 implies that if

$$\int_t^\infty s^{j-k-1}P_j(s)\,ds=O(\phi(t)), \qquad k+1\leqq j\leqq n, \tag{3.20}$$

where $\phi(t)\to 0$ as $t\to\infty$ like some positive power of $1/t$, then (1.5) has solutions $x_l(t)\sim t^l/l!$ $(0\leqq l\leqq n-k-1)$, without any further integrability assumptions on $P_{k+1},\cdots,P_n$, provided that $P_1,\cdots,P_k$ satisfy (2.5). The following corollary shows that this conclusion remains valid even if ϕ decays more slowly than this.

COROLLARY 4. *Suppose that* ϕ *is as in Assumption* A, *and also that* $t^\gamma\phi(t)$ *is eventually nondecreasing for* $\gamma<1$, *and*

$$\int_t^\infty \frac{\phi^2(s)}{s}\,ds=o(\phi(t)). \tag{3.21}$$

Suppose also that (2.5) *and* (3.20) *hold. Then* (1.5) *has solutions* $x_0,\cdots,x_{n-k-1}$ *such that*

$$\left(x_l(t)-\frac{t^l}{l!}\right)^{(r)}=\begin{cases}O(\phi(t)t^{l-r}), & 0\leqq r\leqq n-k-1,\\ O(\phi(t)t^{-n+l+k+1}), & n-k\leqq r\leqq n-1.\end{cases} \tag{3.22}$$

Moreover, if (3.20) *holds with "*O*" replaced by "*o*," then so does* (3.22).

Proof. For each $l=0,1,\cdots,n-k-1$, we apply Theorem 2 with $p(t)=t^l/l!$ and $m=l$. The assumption (3.20) implies (2.4) and also that

$$F_j(t)=\int_t^\infty P_j(s)\,ds=O(\phi(t)t^{-j+k+1}), \qquad k+2\leqq j\leqq n.$$

This implies (2.6) and (2.7) (the latter because of (3.21)), which completes the proof.

A similar argument yields the following related corollary.

COROLLARY 5. *Let ϕ be as in Corollary* 4, *except that* (3.21) *is replaced by*

$$\int_t^\infty \frac{\phi^2(s)}{s}\,ds = O(\phi(t)).$$

Suppose also that (2.5) *holds, and that*

$$\int_t^\infty s^{j-k-1}P_j(s)\,ds = o(\phi(t)), \qquad k+1 \leqq j \leqq n.$$

Then (1.5) *has solutions $x_0, \cdots, x_{n-k-1}$ which satisfy* (3.22) *with "O" replaced by "o."*

Remark 1. Although we have assumed (1.2) throughout, our results are also valid for (1.6), which corresponds to the case where $k=0$, provided that obviously vacuous conditions (i.e., those involving $0 \leqq j \leqq k-1$ and $n-k \leqq r \leqq n-1$) are ignored. To see this, one has only to modify (in fact, simplify) the arguments as follows:

(a) Omit the now vacuous assumptions on the zeros of $Q(\lambda)$.

(b) Let $\mathscr{L}_1$ be the identity operator; i.e., $\mathscr{L}_1 v = v$.

(c) Omit Lemma 1.

Viewed in this way, the present results improve on those in [7] and [8], since we assumed in those papers that $\int^\infty |P_1(t)|\,dt < \infty$, which is stronger that (2.5), (3.5) and (3.18) with $k=0$ and $j=1$.

Remark 2. While preparing this paper the author discovered errors in [7] and [8], caused by his overlooking the need for special treatment of the exceptional cases where (3.2) does not hold. Theorem 2 of [7] requires the additional assumption that

$$\int^\infty \left| \int_t^\infty p_l(\tau)\,d\tau \right| t^{l-2}\,dt < \infty, \qquad 2 \leqq l \leqq n-r-1$$

(the notation here is that of [7]). Our present Corollary 2 (with $k=0$) extends this corrected result. Example 1 of [8] requires the additional assumption that $\alpha \neq 1, \cdots, n-1$. (The notation here is that of [8].) Corollary 1 (with $k=0$) extends this corrected result, and Corollary 2 deals with the excluded cases where $\alpha = 1, \cdots, n-1$.

4. Appendix. Proof of Lemma 1. We assume throughout that $t \geqq t_0 \geqq a$. If

$$V_r(t) = \int_t^\infty s^r v(s)\,ds, \qquad 0 \leqq r \leqq q, \tag{4.1}$$

then (2.26) implies that

$$|V_q(t)| \leqq \psi(t), \tag{4.2}$$

and by writing $s^r v(s) = s^{r-q}(s^q v(s))$, integrating by parts, and again invoking (2.26), we find that

$$|V_r(t)| \leqq 2t^{r-q}\psi(t), \qquad 0 \leqq r \leqq q-1. \tag{4.3}$$

Proof of (i). Now suppose that $\mu > 0$. If $t_0 \leqq s \leqq T$, then integration by parts shows that

$$\int_T^\infty X(s-\tau)\,e^{\lambda(s-\tau)}v(\tau)\,d\tau = X(s-T)\,e^{\lambda(s-T)}V_0(T) + \int_T^\infty V_0(\tau)[X(s-\tau)\,e^{\lambda(s-\tau)}]'\,d\tau.$$

Hence, (4.3) implies that

$$\left| \int_T^\infty X(s-\tau)\,e^{\lambda(s-\tau)}v(\tau)\,d\tau \right| \leqq T^{-q}\psi(T)\hat{X}(T-s)\,e^{\mu(s-T)}, \qquad t_0 \leqq s \leqq T, \tag{4.4}$$

where $\hat{X}$ is a polynomial with nonnegative coefficients determined by X and λ. Setting $s = T = t$ in (4.4) yields (2.29), with $K_1 = \hat{X}(0)$.

To prove (2.30), we consider the integral

$$I(t, T) = \int_t^T s^q f_1(s)\, ds = I_1(t, T) + I_2(t, T), \tag{4.5}$$

where

$$I_1(t, T) = \int_t^T s^q\, ds \int_s^T X(s-\tau)\, e^{\lambda(s-\tau)} v(\tau)\, d\tau \tag{4.6}$$

and

$$I_2(t, T) = \int_t^T s^q\, ds \int_T^\infty X(s-\tau)\, e^{\lambda(s-\tau)} v(\tau)\, d\tau. \tag{4.7}$$

From (4.4) and (4.7),

$$\begin{aligned} |I_2(t, T)| &\leqq \psi(T) \int_t^T \hat{X}(T-s)\, e^{\mu(s-T)}\, ds \\ &< \psi(T) \int_0^\infty \hat{X}(\eta)\, e^{-\mu\eta}\, d\eta; \end{aligned}$$

hence,

$$\lim_{T\to\infty} I_2(t, T) = 0. \tag{4.8}$$

Changing the order of integration in (4.6) yields

$$I_1(t, T) = \int_t^T \Gamma(t, \tau) v(\tau)\, d\tau, \tag{4.9}$$

where

$$\Gamma(t, \tau) = \int_t^\tau s^q X(s-\tau)\, e^{\lambda(s-\tau)}\, ds. \tag{4.10}$$

Repeated integration by parts shows that

$$\Gamma(t, \tau) = \sum_{r=0}^{q} [X_r(0)\tau^r - t^r X_r(t-\tau)\, e^{\lambda(t-\tau)}], \tag{4.11}$$

where $X_0, \cdots, X_q$ are polynomials determined by X and λ. Substituting (4.11) into (4.9), we obtain $I(t, T)$ in terms of integrals which converge as $T \to \infty$; therefore, (4.5), (4.8) and (4.11) imply that the integral $f_2(t) = I(t, \infty)$ (see (2.28)) converges, and that

$$f_2(t) = \sum_{r=0}^{q} X_r(0)\, V_r(t) - \sum_{r=0}^{q} t^r \int_t^\infty X_r(t-\tau)\, e^{\lambda(t-\tau)} v(\tau)\, d\tau \tag{4.12}$$

(see (4.1)). Replacing X by X_r in (4.4) and letting $s = T = t$ shows that

$$\left| \int_t^\infty X_r(t-\tau)\, e^{\lambda(t-\tau)} v(\tau)\, d\tau \right| \leqq K_{1r} t^{-q} \psi(t),$$

where K_{1r} is a constant determined by λ and X_r (and therefore by X). This, (4.2), (4.3) and (4.12) imply (2.30) for suitable K_2, a constant determined by X and λ. This proves (i).

Proof of (ii). Now suppose that $\mu < 0$. Integrating (2.33) by parts yields

$$f_3(t) = V_0(t_0)X(t-t_0)\, e^{\lambda(t-t_0)} - X(0)\, V_0(t) + \int_{t_0}^{t} V_0(\tau)[X(t-\tau)\, e^{\lambda(t-\tau)}]'\, d\tau.$$

Therefore, from (4.3) with $r = 0$,

$$\begin{aligned} |f_3(t)| \leqq 2t_0^{-q}\psi(t_0)|X(t-t_0)|\, e^{\mu(t-t_0)} + 2|X(0)|\psi(t)t^{-q} \\ + 2\int_{t_0}^{t} \psi(\tau)\tau^{-q}\tilde{X}(t-\tau)\, e^{\mu(t-\tau)}\, d\tau \end{aligned} \tag{4.13}$$

where $\tilde{X}$ is a polynomial with positive coefficients determined by X and λ. Now (2.32) and our assumptions on α and ϕ imply that

$$\psi(\tau)\tau^{-q} \leqq \psi_1(t_0)\, e^{\alpha(t-\tau)}\phi(t)t^{-q}, \qquad t \geqq \tau \geqq t_0.$$

This and (4.13) imply (2.34), with

$$K_3 = 2\left[|X(0)| + \sup_{\xi \geqq 0}|X(\xi)|\, e^{(\alpha+\mu)\xi} + \int_0^{\infty} \tilde{X}(\eta)\, e^{(\alpha+\mu)\eta}\, d\eta\right],$$

which is finite, because of (2.31).

To see that (2.37) implies (2.38), observe from (2.33) that if $t_0 \leqq t_1 \leqq t$, then

$$f_3(t) = \int_{t_0}^{t_1} X(t-\tau)\, e^{\lambda(t-\tau)}v(\tau)\, d\tau + \int_{t_1}^{t} X(t-\tau)\, e^{\lambda(t-\tau)}v(\tau)\, d\tau. \tag{4.14}$$

For fixed t_1,

$$\left|\int_{t_0}^{t_1} X(t-\tau)\, e^{\lambda(t-\tau)}v(\tau)\, d\tau\right| \leqq M(t_0, t_1)\, Y(t)\, e^{\mu t}, \tag{4.15}$$

where

$$M(t_0, t_1) = \max\{|v(\tau)|\,|\, t_0 \leqq \tau \leqq t_1\}$$

and Y is a polynomial, while

$$\left|\int_{t_1}^{t} X(t-\tau)\, e^{\lambda(t-\tau)}v(\tau)\, d\tau\right| \leqq K_3\psi_1(t_1)t^{-q}\phi(t), \qquad t \geqq t_1. \tag{4.16}$$

(Since this integral has the same form as $f_3(t)$ in (2.33), with t_0 replaced by t_1, (4.16) is obtained as was (2.34).)

Our assumptions on ϕ and α imply that

$$Y(t)\, e^{\mu t} = o(t^{-q}\phi(t)).$$

Therefore, (4.14), (4.15) and (4.16) imply that

$$\overline{\lim_{t\to\infty}}\, (t^{-q}\phi(t))^{-1}|f_3(t)| \leqq K_3\psi_1(t_1), \qquad t_1 \geqq t_0.$$

Since (2.32) and (2.37) imply that $\lim_{t_1\to\infty}\psi_1(t_1) = 0$, we now see that (2.37) implies (2.38).

For reference below, we observe here that

$$\lim_{t\to\infty} t^q f_3(t) = 0 \tag{4.17}$$

in any case. This is obvious from (2.34) if $\phi(t)=o(1)$. If $\lim_{t\to\infty}\phi(t)>0$, then (2.26) implies (2.37), which implies (2.38) and, therefore, (4.17).

To prove (2.36), consider the integral

$$J(t,T)=\int_t^T s^q f_3(s)\,ds=\int_t^T s^q\,ds\int_{t_0}^s X(s-\tau)\,e^{\lambda(s-\tau)}v(\tau)\,d\tau.$$

Reversing the order of integration yields

$$J(t,T)=\int_{t_0}^t v(\tau)\,d\tau\int_t^T s^qX(s-\tau)\,e^{\lambda(s-\tau)}\,ds+\int_t^T v(\tau)\,d\tau\int_\tau^T s^qX(s-\tau)\,e^{\lambda(s-\tau)}\,ds.$$

Manipulating the limits of integration here shows that

$$J(t,T)=H(t)-H(T).$$

where

$$H(t)=\int_{t_0}^t \Gamma(t,\tau)v(\tau)\,d\tau\quad\text{(see (4.10)).}$$

Therefore, from (4.11),

$$H(t)=\sum_{r=0}^q X_r(0)\int_{t_0}^t \tau^r v(\tau)\,d\tau-\sum_{r=0}^q t^r\int_{t_0}^t X_r(t-\tau)\,e^{\lambda(t-\tau)}v(\tau)\,d\tau. \tag{4.18}$$

The integrals on the right converge as $t\to\infty$. Moreover, since the integrals in the second sum are of the same form as $f_3(t)$ (see (2.33)), (4.17) implies that this sum approaches zero as $t\to\infty$; hence,

$$H(\infty)=\lim_{T\to\infty}H(T)=\sum_{r=0}^q X_r(0)\int_{t_0}^\infty \tau^r v(\tau)\,d\tau. \tag{4.19}$$

Since $f_4(t)=H(\infty)-H(t)$, we now see from (4.1), (4.18) and (4.19) that

$$f_4(t)=\sum_{r=0}^q X_r(0)V_r(t)+\sum_{r=0}^q t^r\int_{t_0}^t X_r(t-\tau)\,e^{\lambda(t-\tau)}v(\tau)\,d\tau. \tag{4.20}$$

Recalling (2.32), (4.2), (4.3) and, again, that the integrals in the second sum here are of the same form as $f_3(t)$, and therefore satisfy an inequality like (2.34), we see from (4.20) that (2.36) holds for a suitable constant K_4 which is ultimately determined by X and λ.

Finally, suppose that (2.37) holds. Then obviously $V_r(t)=o(\phi(t))$, from (4.2) and (4.3). Moreover,

$$\int_{t_0}^t X_r(t-\tau)\,e^{\lambda(t-\tau)}v(\tau)\,d\tau=o(t^{-q}\phi(t)),$$

as can be seen by replacing X with X_r in (2.33) and applying the argument that led to (2.38). Now (4.20) implies (2.39), which completes the proof of Lemma 1.

REFERENCES

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

[2] ———, *Asymptotic integration of ordinary differential equations*, this Journal, 14 (1983), pp. 772-779.

[3] T. W. PREVATT, *Application of exponential dichotomies to asymptotic integration and the spectral theory of ordinary differential equations*, J. Differential Equations, 17 (1975), pp. 444-460.

[4] J. ŠIMŠA, *Asymptotic integration of perturbed linear differential equations under conditions involving ordinary integral convergence*, this Journal, 15 (1984), pp. 116–123.

[5] ———, *The second order differential equation with oscillatory coefficient*, Arch. Math. (Brno), 18 (1982), pp. 95–100.

[6] ———, *The condition of ordinary integral convergence in the asymptotic theory of linear differential equations with almost constant coefficients*, this Journal, 16 (1985), pp. 757–769.

[7] W. F. TRENCH, *Asymptotic integration of linear differential equations subject to integral smallness conditions involving ordinary convergence*, this Journal, 7 (1976), pp. 213–221.

[8] ———, *Asymptotic integration of linear differential equations subject to mild integral conditions*, this Journal, 15 (1984), pp. 932–942.

[9] ———, *Linear perturbations of a constant coefficient differential equation subject to mild integral smallness conditions*, Czechoslovak Math J., 36 (1986), pp. 623–633.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

014

AN EXTENSION OF A THEOREM OF PERRON*

JAROMÍR ŠIMŠA†

Abstract. We give sufficient conditions for a linear differential equation of order n to have a solution whose logarithmic derivative is asymptotically in a small neighbourhood of a given constant. This asymptotic behavior is specified by means of a general comparison function φ. An appropriate formulation of the growth properties of φ is introduced. The differential equation is regarded as a perturbation of some constant coefficient equation. Our smallness conditions permit conditional convergence of some corresponding improper integrals.

Key words. ordinary linear differential equations, linear perturbations, asymptotic integration

AMS(MOS) subject classification. 34E99

1. Introduction and main theorem. We study the behavior for large t of the solutions of a scalar differential equation

$$x^{(n)}+[a_1+p_1(t)]x^{(n-1)}+\cdots+[a_n+p_n(t)]x=0 \qquad (t\geqq 0,\ n\geqq 2), \tag{1}$$

where a_k are constants and p_k are continuous on $[0,\infty)$. We assume throughout that a_k, p_k and x are real- or complex-valued. The symbols $\langle\langle O\rangle\rangle$ and $\langle\langle o\rangle\rangle$ refer to behavior as $t\to\infty$.

Perron [4] showed that if

$$p_k(t)=o(1) \qquad (1\leqq k\leqq n), \tag{2}$$

then (1) has a solution x_0 satisfying

$$\frac{x_0^{(k)}(t)}{x_0(t)}=\lambda_0^k+o(1) \qquad (1\leqq k\leqq n-1), \tag{3}$$

provided that λ_0 is a simple root of

$$\lambda^n+a_1\lambda^{n-1}+\cdots+a_{n-1}\lambda+a_n=0 \tag{4}$$

and $\operatorname{Re}\lambda_j\neq\operatorname{Re}\lambda_0$, for any other root λ_j of (4). Hartman and Wintner [3] have showed that Perron's conclusion remains valid if (2) is relaxed to

$$\sup_{s\geqq t}(1+s-t)^{-1}\int_t^s|p_k(r)|\,dr=o(1) \qquad (1\leqq k\leqq n) \tag{5}$$

(see also [1, Thm. 17.4]).

In this paper, we obtain another extension of Perron's Theorem. Namely, we give sufficient conditions for (1) to have a solution x_0 satisfying

$$\frac{x_0^{(k)}(t)}{x_0(t)}=\lambda_0^k+o(\varphi(t)) \qquad (1\leqq k\leqq n-1), \tag{6}$$

where the function φ is positive, continuous and nonincreasing on $[T,\infty)$, for some real T. Of course, due to the behavior of φ, we impose a new restriction on the roots of (4). In the case $\varphi=1$, when (6) becomes (3), our result is stronger than that of Hartman and Wintner mentioned above, because of the considerable weakening of (5).

* Received by the editors March 20, 1986; accepted for publication (in revised form) August 6, 1986.
† Department of Mathematics, J. E. Purkyně University, Janáčkovo nám. 2a, 662 95 Brno, Czechoslovakia.

Before we state the main theorem, we introduce the following notation. Assuming the function φ to be as above, we define

$$m_\varphi = \sup\{\alpha: e^{\alpha t}\varphi(t) \text{ is nonincreasing on } [T_\alpha, \infty) \text{ for some } T_\alpha \geqq T\} \tag{7}$$

and

$$M_\varphi = \inf\{\alpha: e^{\alpha t}\varphi(t) \text{ is nondecreasing on } [T_\alpha, \infty) \text{ for some } T_\alpha \geqq T\}, \tag{8}$$

where M_φ is meant to be ∞, if the set in (8) is empty. Obviously, it holds that $0 \leqq m_\varphi \leqq M_\varphi \leqq \infty$.

THEOREM 1. *Let λ_0 be a root of* (4) *such that*

$$\operatorname{Re}(\lambda_0 - \lambda_j) \notin [m_\varphi, M_\varphi] \tag{9}$$

for any zero λ_j of the polynomial

$$Q(\lambda) = (\lambda - \lambda_0)^{-1}(\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n). \tag{10}$$

(*Thus λ_0 need not be simple if $m_\varphi > 0$.*) *Suppose that*

$$\sup_{s \geqq t} (1+s-t)^{-1} \int_t^s |p_1(r)|\varphi(r)\, dr = o(\varphi(t)), \tag{11}$$

$$\sup_{s \geqq t} (1+s-t)^{-1} \left| \int_t^s p_k(r)\, dr \right| = o(1) \qquad (2 \leqq k \leqq n) \tag{12}$$

and

$$\sup_{s \geqq t} (1+s-t)^{-1} \left| \int_t^s f(r)\, dr \right| = o(\varphi(t)), \tag{13}$$

where

$$f(t) = \sum_{k=1}^{n} \lambda_0^{n-k} p_k(t) \qquad (0 \leqq t < \infty). \tag{14}$$

Then (1) *has a solution x_0 satisfying* (6). *This conclusion remains valid if* (12) *is relaxed to*

$$\sup_{s \geqq t} (1+s-t)^{-1} \left| \int_t^s (p_k(r) - q_k(r))\, dr \right| = o(1) \qquad (2 \leqq k \leqq n), \tag{15}$$

where q_k are some continuous functions satisfying

$$\sup_{s \geqq t} (1+s-t)^{-1} \int_t^s |q_k(r)|\varphi(r)\, dr = o(\varphi(t)) \qquad (2 \leqq k \leqq n). \tag{16}$$

Remark 1. A mild condition like (12) previously appeared in Hartman's work [2]. The result there is that if λ_0 and p are real-valued, $\lambda_0 \neq 0$ and if p is continuous on $[0, \infty)$, then the condition

$$\sup_{s \geqq t} (1+s-t)^{-1} \left| \int_t^s p(r)\, dr \right| = o(1)$$

is necessary and sufficient for the equation $x'' - (\lambda_0^2 + p(t))x = 0$ to have a solution x_0 satisfying (3) with $n = 2$ (see also [1, Exer. 17.3]).

Remark 2. The statement of Theorem 1 can be simplified if $m_\varphi > 0$ or, more generally, if $\varphi \in L_1$ and

$$\int_t^\infty \varphi(s)\, ds = O(\varphi(t)). \tag{17}$$

Then (6) is equivalent to

$$x_0^{(k)}(t) = [c\lambda_0^k + o(\varphi(t))]\, e^{\lambda_0 t} \qquad (0 \leqq k \leqq n-1), \tag{18}$$

where $c \neq 0$ is a constant, while (11), (13) and (16) mean that the integrals

$$\int_t^\infty |p_1(s)|\varphi(s)\, ds, \quad \int_t^\infty f(s)\, ds, \quad \int_t^\infty |q_k(s)|\varphi(s)\, ds \qquad (2 \leqq k \leqq n) \tag{19}$$

exist and are $o(\varphi(t))$. The proof of these assertions is given in the end of § 2.

2. Preliminary lemmas. The following lemmas will be used to prove Theorem 1 in § 3 and its corollaries in § 4.

LEMMA 1. *Suppose that h is in $C[T, \infty)$ and φ is positive and nonincreasing on $[T, \infty)$. If*

$$\sup_{s \geqq t} (1+s-t)^{-1} \left| \int_t^s h(r)\, dr \right| \leqq K_0 \psi(t) \qquad (T \leqq t < \infty) \tag{20}$$

holds with $K_0 = 1$, then the integral $\int^\infty h(t)\, e^{\gamma t}\, dt$ converges and satisfies

$$\left| \int_t^\infty h(s)\, e^{\gamma s}\, ds \right| \leqq K_1\, e^{\gamma t} \psi(t) \qquad (T \leqq t < \infty), \tag{21}$$

where the constant K_1 depends on γ only, for any $\gamma < 0$. Conversely if the integral $\int^\infty h(t)\, e^{\gamma t}\, dt$ converges and satisfies (21) *with $K_1 = 1$, for some $\gamma \leqq 0$, then* (20) *holds with a constant K_0 which depends on γ only.*

Proof. Denote

$$H(s) = \int_t^s h(r)\, dr$$

and suppose that $|H(s)| \leqq (1+s-t)\psi(t)$ for any $s \geqq t$. If $\gamma < 0$, then $H(s)e^{\gamma s} \to 0$ $(s \to \infty)$ and

$$\begin{aligned} \int_t^\infty |H(s)|\, e^{\gamma s}\, ds &\leqq \psi(t) \int_t^\infty (1+s-t)\, e^{\gamma s}\, ds \\ &= \psi(t)\, e^{\gamma t} \int_0^\infty (1+r) e^{\gamma r}\, dr. \end{aligned}$$

Consequently, we can let $r \to \infty$ in

$$\int_t^r h(s)\, e^{\gamma s}\, ds = (H(s)\, e^{\gamma s})\Big|_t^r - \gamma \int_t^r H(s)\, e^{\gamma s}\, ds$$

and conclude that $\int^\infty h(t)\, e^{\gamma t}\, dt$ converges and (21) holds if

$$K_1 = |\gamma| \int_0^\infty (1+r)\, e^{\gamma r}\, dr.$$

Conversely, denote

$$H_1(t) = \int_t^\infty h(s)\, e^{\gamma s}\, ds \quad \text{and} \quad H_2(t) = \sup_{s \geqq t} |H_1(s)|\, e^{-\gamma s},$$

and suppose that $H_2(t) \leqq \psi(t)$ for some $\gamma \leqq 0$. Integration by parts yields

$$\left| \int_t^s h(r)\, dr \right| = \left| (-H_1(r)\, e^{-\gamma r})|_t^s - \gamma \int_t^s H_1(r)\, e^{-\gamma r}\, dr \right|$$

$$\leqq 2H_2(t) + |\gamma| \int_t^s H_2(r)\, dr \leqq [2 + |\gamma|(s-t)]\psi(t)$$

for any $s \geqq t$. Consequently, (20) holds if K_0 is an upper bound of $(1+r)^{-1}(2+|\gamma|r)$ for $r \geqq 0$. This completes the proof of Lemma 1.

LEMMA 2. *Lemma* 1 *remains valid if* (20) *and* (21) *are replaced by*

$$\sup_{s \geqq t} [1 + \log(s/t)]^{-1} \left| \int_t^s h(r)\, dr \right| \leqq K_0 \psi(t) \qquad (T \leqq t < \infty)$$

and

$$\left| \int_t^\infty h(s) s^\gamma\, ds \right| \leqq K_1 t^\gamma \psi(t) \qquad (T \leqq t < \infty),$$

respectively.

Proof. Substituting $u = \log r$ yields

$$\int_t^s h(r)\, dr = \int_{\log t}^{\log s} h(e^u)\, e^u\, du, \quad \text{and}$$

$$\int_t^\infty h(r)\, r^\gamma\, dr = \int_{\log t}^\infty h(u)\, e^{(\gamma+1)u}\, du.$$

Thus Lemma 2 follows from Lemma 1, with $\psi(t)$ and $h(t)$ replaced by $\psi(e^t)$ and $e^t h(t)$, respectively.

LEMMA 3. *Assume that* σ *and* ψ *are nonnegative functions in* $C[t_0, \infty)$, σ *is nonincreasing,* B *is a polynomial with nonnegative coefficients and* $\gamma \neq 0$ *is a real constant. Define*

$$w(t) = \int_t^\infty e^{-|\gamma|s} B(s)\, ds \qquad (t \geqq 0). \tag{22}$$

If $\gamma > 0$ *and* $e^{\gamma t}\psi(t)$ *is nonincreasing on* $[t_0, \infty)$, *then*

$$\int_t^\infty B(s-t)\sigma(s)\psi(s)\, ds \leqq w(0)\sigma(t)\psi(t) \qquad (t \geqq t_0). \tag{23}$$

If $\gamma < 0$ *and* $e^{\gamma t}\psi(t)$ *is nondecreasing on* $[t_0, \infty)$, *then*

$$\int_{t_0}^t B(t-s)\sigma(s)\psi(s)\, ds \leqq [w(t-t_1)\sigma(t_0) + w(0)\sigma(t_1)]\psi(t) \qquad (t \geqq t_0), \tag{24}$$

where $t_1 = (t_0 + t)/2$.

Proof. If $\gamma > 0$ and $\psi(s) \leqq \psi(t)\, e^{\gamma(t-s)}$ for any $s \geqq t$, then the integral in (23) does not exceed

$$\sigma(t)\psi(t) \int_t^\infty e^{\gamma(t-s)} B(s-t)\, ds = w(0)\sigma(t)\psi(t).$$

If $\gamma<0$ and $\psi(s)\leqq\psi(t)\,e^{\gamma(t-s)}$, for any s in $[t_0, t]$, then the integral in (24) does not exceed

$$\psi(t)\left[\sigma(t_0)\int_{t_0}^{t_1} e^{\gamma(t-s)}B(t-s)\,ds+\sigma(t_1)\int_{t_1}^{t} e^{\gamma(t-s)}B(t-s)\,ds\right],$$

where $t_1=(t_0+t)/2$. Thus (24) holds, because

$$\int_{t_3}^{t_2} e^{\gamma(t-s)}B(t-s)\,ds\leqq\int_{-\infty}^{t_2} e^{\gamma(t-s)}B(t-s)\,ds=\hat{w}(t-t_2),$$

whenever $t_3\leqq t_2\leqq t$. This completes the proof of Lemma 3.

Before we state the last lemma of this section, we introduce a factor smallness condition as in [6, p. 763].

DEFINITION 1. Let $\sigma(t_0, t)$ be a real function defined for all t_0 and t, $T\leqq t_0<\infty$, which is nonincreasing in t and $\sigma(t_0, t)\to 0$ as $t\to\infty$ for any $t_0\geqq T$. If in addition $\sigma(t_0, t_0)\to 0$ as $t_0\to\infty$, then we sat that σ is of type (*).

LEMMA 4. *Assume that A is a polynomial, ψ_1, ψ_2 and p are in $C[T,\infty)$, ψ_1 and ψ_2 are positive, the integral $\int^\infty p(t)\,dt$ converges (perhaps conditionally) and satisfies*

$$\sigma(t)=\sup_{s\geqq t}(\psi_1(s))^{-1}\left|\int_s^\infty p(r)\,dr\right|\to 0\qquad(t\to\infty). \tag{25}$$

Assume also K is a function in $C^1[t_0,\infty)$ satisfying $|K^{(j)}(t)|\leqq K_0\psi_2(t)$, where K_0 is a constant, $t\geqq t_0$ and $j=0, 1$, for some $t_0\geqq T$. Denote $\psi=\psi_1\psi_2$.

(i) *If $e^{\gamma t}\psi(t)$ is nonincreasing on $[t_0,\infty)$, for some $\gamma>0$, then the integral $\int^\infty A(t-s)K(s)p(s)\,ds$ converges and satisfies*

$$\left|\int_t^\infty A(t-s)K(s)p(s)\,ds\right|\leqq K_0C\sigma(t)\psi(t)\qquad(t\geqq t_0), \tag{26}$$

where C is a constant which depends only on A and γ.

(ii) *If $e^{\gamma t}\psi(t)$ is nondecreasing on $[t_0,\infty)$ for some $\gamma<0$, then*

$$\left|\int_{t_0}^t A(t-s)K(s)p(s)\,ds\right|\leqq K_0\sigma_1(t_0, t)\psi(t)\quad(t\geqq t_0), \tag{27}$$

where σ_1 is of type (*) and depends only on σ, A and γ.

Proof. Our assumptions on K imply that

$$|(A(t-s)K(s))'|\leqq K_0B(|t-s|)\psi_2(s)\qquad\left(s, t\geqq t_0, '=\frac{d}{ds}\right), \tag{28}$$

where B is a polynomial with nonnegative coefficients which satisfies $B(s)\geqq|A(s)|+|A'(s)|$ for any $s\geqq 0$. From (25),

$$|P(t)|\leqq\sigma(t)\psi_1(t)\quad\text{where } P(t)=\int_t^\infty p(s)\,ds \text{ and } t\geqq t_0. \tag{29}$$

(i) Integration by parts yields

$$\int_t^\infty A(t-s)K(s)p(s)\,ds=A(0)K(t)P(t)+\int_t^\infty(A(t-s)K(s))'P(s)\,ds. \tag{30}$$

Lemma 3, (28) and (29) imply that the integral on the right side of (30) converges absolutely, and that (26) holds with $C=|A(0)|+w(0)$, where w is as in (22).

(ii) Integration by parts yields

$$(31)\quad \int_{t_0}^{t} A(t-s)K(s)p(s)\,ds = -A(t-s)K(s)P(s)\big|_{t_0}^{t} + \int_{t_0}^{t} (A(t-s)K(s))'P(s)\,ds.$$

If $t_0 \leqq s \leqq t$, then our assumptions on K and γ, (28) and (29) imply that

$$(32)\quad \begin{aligned} |A(t-t_0)K(t_0)P(t_0)| &\leqq K_0 B(t-t_0)\sigma(t_0)\psi(t_0) \\ &\leqq K_0 B(t-t_0)\sigma(t_0)\, e^{\gamma(t-t_0)}\psi(t) \end{aligned}$$

and

$$(33)\quad |(A(t-s)K(s))'P(s)| \leqq K_0 B(t-s)\sigma(s)\psi(s).$$

Now Lemma 3, (29) and (31)-(33) imply that (27) holds, with σ_1 given by

$$\begin{aligned} \sigma_1(t_0, t) = \sigma(t_0)[w(t-t_1) &+ \sup_{s \geqq t-t_0} e^{\gamma s} B(s)] \\ &+ \sigma(t_1)w(0) + \sigma(t)|A(0)|, \end{aligned}$$

where $t_1 = (t+t_0)/2$ and w is as in (22). Obviously, σ_1 is of type (*) and depends only on σ, A and γ. This completes the proof of Lemma 4.

We finish this section by proving the assertions of Remark 2 in § 1. First we show that $m_\varphi > 0$ implies (17). Indeed, if $e^{\gamma t}\varphi(t)$ is nonincreasing on $[T\gamma, \infty)$ for some $\gamma > 0$ (see (7)), then

$$\int_t^\infty \varphi(s)\,ds \leqq \int_t^\infty \varphi(t)\, e^{\gamma(t-s)}\,ds = \gamma^{-1}\varphi(t) \qquad (t \geqq T\gamma).$$

To prove (18), we need only to show that

$$(34)\quad x_0(t) = [c + o(\varphi(t))]\, e^{\lambda_0 t},$$

because (6) and (34) easily imply (18). Integrating (6) with $k=1$, we obtain

$$x_0(t) = x_0(t_0) \exp\left[\lambda_0(t-t_0) + \int_{t_0}^{t} o(\varphi(s))\,ds\right],$$

which proves (34), because of (17). Finally, we verify that (11), (13) and (16) imply the assertion on the integrals (19). (The converse is true by Lemma 1 with $\gamma = 0$, without supposing (17).) It suffices to show that the integral $\int^\infty h(t)\,dt$ converges and satisfies

$$(35)\quad \int_t^\infty h(s)\,ds = o(\varphi(t)),$$

provided that (17) holds and

$$(36)\quad \sup_{s \geqq t} (1+s-t)^{-1} \left| \int_t^s h(r)\,dr \right| = o(\varphi(t)).$$

To see this, observe that Lemma 1 and (36) imply that

$$(37)\quad H_1(t) = \int_t^\infty e^{\gamma s} h(s)\,ds = o(e^{\gamma t}\varphi(t))$$

if γ is a negative constant. From (17) and (37), we can let $r \to \infty$ in

$$\int_t^r h(s)\,ds = -H_1(s)\, e^{-\gamma s}\big|_t^r - \gamma \int_t^r H_1(s)\, e^{-\gamma s}\,ds$$

and conclude that (35) holds.

3. Asymptotic integration. This section contains the proof of Theorem 1, stated in § 1. The method of asymptotic integration is similar to that of [6], and we do not repeat some arguments given there.

We prove Theorem 1 assuming the function φ to satisfy

$$\varphi(t) = o(1). \tag{38}$$

(The case that $\lim_{t\to\infty} \varphi(t) > 0$ is discussed in the end of this section.) Furthermore, it suffices to assume that $\lambda_0 = 0$, otherwise $x \exp(-\lambda_0 t)$ is introduced as a new dependent variable in (1). However, due to this change, (12) may fail to hold, because (11) does not imply (12) with $k = 1$. Therefore, we give the proof with (12) replaced by the weaker assumption (15).

Let λ_j $(1 \leqq j \leqq L)$ be all distinct zeros of the polynomial $Q(\lambda) = \lambda^{n-1} + a_1\lambda^{n-2} + \cdots + a_{n-1}$ (see (10) with $\lambda_0 = a_n = 0$), and let $\operatorname{Re}\lambda_1 \leqq \operatorname{Re}\lambda_2 \leqq \cdots \leqq \operatorname{Re}\lambda_L$. From (9) with $\lambda_0 = 0$, there exists an integer $N (0 \leqq N \leqq L)$ such that

$$\operatorname{Re}\lambda_j + M_\varphi < 0 \quad (0 < j \leqq N), \qquad \operatorname{Re}\lambda_j + m_\varphi > 0 \quad (N < j \leqq L). \tag{39}$$

Now we introduce the new dependent variable u in (1) by $x' = ux$. Then

$$x^{(k)} = (u^{(k-1)} + g_k[u])x \qquad (1 \leqq k \leqq n), \tag{40}$$

where $g_k[u]$ are nonlinear expressions which have been described in detail in [6, p. 759]. Substituting (40) into (1) with $a_n = 0$, we obtain

$$Q(D)u = -p_n - R[u] \qquad \left(D = \frac{d}{dt}\right) \tag{41}$$

where

$$R[u] = \sum_{k=1}^{n-1} p_k u^{(n-k-1)} + \sum_{k=0}^{n-1} a_k g_{n-k}[u] + \sum_{k=1}^{n-1} p_k g_{n-k}[u] \qquad (a_0 = 1). \tag{42}$$

What follows is the standard variation of constants, applied to (41). Let $U(t)$ be the unique solution of $Q(D)u = 0$ that satisfies $D^i U(0) = 0 (0 \leqq i \leqq n-3)$ and $D^{n-2}U(0) = 1$. Then there exist solutions U_1 and U_2 of $Q(D)u = 0$ such that $U = U_1 + U_2$ and

$$D^i U_1(t) = \sum_{j=1}^{N} A_{ji}(t)\, e^{\lambda_j t}, \qquad D^i U_2(t) = \sum_{j=N+1}^{L} A_{ji}(t)\, e^{\lambda_j t} \tag{43}$$

where A_{ji} are polynomials. If we define

$$\mathscr{L}h(t) = \int_{t_0}^{t} U_1(t-s)h(s)\, ds - \int_{t}^{\infty} U_2(t-s)h(s)\, ds \tag{44}$$

for any continuous h and some $t_0 \geqq T$, then routine computations show that

$$D^i\mathscr{L}h(t) = \int_{t_0}^{t} D^i U_1(t-s)h(s)\, ds - \int_{t}^{\infty} D^i U_2(t-s)h(s)\, ds \qquad (0 \leqq i \leqq n-2) \tag{45}$$

and $Q(D)\mathscr{L}h = h$, provided that the improper integrals in (45) converge. Consequently, (41) can be converted into an integral equation

$$u = -\mathscr{L}p_n - \mathscr{L}R[u]. \tag{46}$$

The solution of (46) will be found as a fixed point (function) of the mapping

$$\mathscr{T}u = -\mathscr{L}p_n - \mathscr{L}R[u] \tag{47}$$

in the set $\mathcal{B}(t_0)$ of all functions u in $C^{n-2}[t_0, \infty)$ satisfying

$$D^k u(t) = O(\varphi(t)) \qquad (0 \leqq k \leqq n-2) \tag{48}$$

which is a Banach space with norm

$$\|u\| = \sup_{t \geqq t_0} \max_{0 \leqq k \leqq n-2} (\varphi(t))^{-1} |D^k u(t)|. \tag{49}$$

We will show that $\mathcal{T}$ is a contraction mapping of the closed sphere

$$\mathcal{S}(t_0) = \{u \in \mathcal{B}(t_0) : \|u\| \leqq 1\}$$

into itself if t_0 is sufficiently large.

In the following, assume that $t \geqq t_0$. From (49),

$$|D^k u(t)| \leqq \|u\| \varphi(t) \qquad (0 \leqq k \leqq n-2), \tag{50}$$

for any u in $\mathcal{B}(t_0)$. Moreover, there exists a universal constant C such that the estimates

$$|g_k[\tilde{u}](t) - g_k[\tilde{\tilde{u}}](t)| \leqq C \|\tilde{u} - \tilde{\tilde{u}}\| \varphi^2(t) \qquad (1 \leqq k \leqq n) \tag{51}$$

and

$$|g_k'[\tilde{u}](t) - g_k'[\tilde{\tilde{u}}](t)| \leqq C \|\tilde{u} - \tilde{\tilde{u}}\| \varphi^2(t) \qquad (1 \leqq k \leqq n-1) \tag{52}$$

hold for any functions $\tilde{u}$, $\tilde{\tilde{u}}$ in $\mathcal{S}(t_0)$. The last can be proved in the same way as in [6, p. 764], where the case $\varphi(t) = t^{-q}$ was considered.

We assume henceforth that t_0 is so large that $\exp[-(\operatorname{Re} \lambda_j + \gamma)t]\psi(t)$ is nondecreasing on $[t_0, \infty)$ if $1 \leqq j \leqq N$, and that $\exp[-(\operatorname{Re} \lambda_j - \gamma)t]\psi(t)$ is nonincreasing on $[t_0, \infty)$ if $N+1 \leqq j \leqq L$, for some positive γ (see (7), (8) and (39)).

LEMMA 5. *The functions $\mathcal{L}p_n$ and $\mathcal{L}R[u]$ exist and are in $C^{n-1}[t_0, \infty)$ for any u in $\mathcal{S}(t_0)$. Moreover, there exists a function σ of type* (*) *(see Definition* 1 *in* § 2*) such that the estimates*

$$|\mathcal{D}^i \mathcal{L}p_n(t)| \leqq \sigma(t_0, t)\varphi(t) \qquad (0 \leqq i \leqq n-2), \tag{53}$$

and

$$|D^i \mathcal{L}R[\tilde{u}](t) - D^i \mathcal{L}R[\tilde{\tilde{u}}](t)| \leqq \|\tilde{u} - \tilde{\tilde{u}}\| \sigma(t_0, t)\varphi(t) \qquad (0 \leqq i \leqq n-2) \tag{54}$$

hold for any $\tilde{u}$, $\tilde{\tilde{u}}$ in $\mathcal{S}(t_0)$.

Proof. Lemma 1, (11), (13) with $f = p_n$ (see (14) with $\lambda_0 = 0$), (15) and (16) imply that

$$\int_t^\infty |p_1(s)|\varphi(s)\, e^{cs}\, ds = o(\varphi(t)\, e^{ct}), \tag{55}$$

$$\int_t^\infty (p_k(s) - q_k(s))\, e^{cs}\, ds = o(e^{ct}) \qquad (2 \leqq k \leqq n), \tag{56}$$

$$\int_t^\infty |q_k(s)|\varphi(s)\, e^{cs}\, ds = o(\varphi(t)\, e^{ct}) \qquad (2 \leqq k \leqq n), \tag{57}$$

and

$$\int_t^\infty p_n(s)\, e^{cs}\, ds = o(\varphi(t)\, e^{ct}), \tag{58}$$

where c is a negative constant. (The integrals in (56) and (58) may converge conditionally.) From (43) and (45),

$$D^i\mathscr{L}h(t)=\sum_{j=1}^{N} e^{\lambda_j t}\int_{t_0}^{t} e^{-\lambda_j s}A_{ji}(t-s)h(s)\,ds - \sum_{j=N+1}^{L} e^{\lambda_j t}\int_{t}^{\infty} e^{-\lambda_j s}A_{ji}(t-s)h(s)\,ds \qquad (0\leqq i\leqq n-2). \tag{59}$$

To prove (53) and (54), we estimate each of the integrals in (59), for $h=p_n$ and for $h=R[\tilde{u}]-R[\tilde{\tilde{u}}]$, using Lemmas 3 and 4. To avoid unnecessary repetition, we note before that the estimates of $|K^{(j)}(t)|$ and the integrability condition on p follow from (50)-(52) and (55)-(58). For the monotonic property of ψ see the sentence preceding the statement of Lemma 5.

First observe that (53) follows from (59) with $h=p_n$ and from Lemma 4, with $p(t)=p_n(t)\,e^{ct}$, $\psi_1(t)=e^{ct}\varphi(t)$, $\psi_2(t)=\exp[-(\operatorname{Re}\lambda_j+c)t]$, $K(t)=\exp[-(\lambda_j+c)t]$ and $K_0=1+|\lambda_j+c|$, for each $j=1,2,\cdots,L$. The case $n=R[\tilde{u}]-R[\tilde{\tilde{u}}]$ is more complicated. According to (42) and our integrability conditions (55)-(58), we can write

$$\begin{aligned} R[\tilde{u}]-R[\tilde{\tilde{u}}]=&\sum_{k=2}^{n-1}(p_k-q_k)u^{(n-k-1)}+\sum_{k=2}^{n-1}q_k u^{(n-k-1)}\\ &+p_1u^{(n-2)}+\sum_{k=0}^{n-1}a_k g_{n-k}[\tilde{u},\tilde{\tilde{u}}]+p_1g_{n-1}[\tilde{u},\tilde{\tilde{u}}]\\ &+\sum_{k=2}^{n-1}(p_k-q_k)g_{n-k}[\tilde{u},\tilde{\tilde{u}}]+\sum_{k=2}^{n-1}q_k g_{n-k}[\tilde{u},\tilde{\tilde{u}}], \end{aligned} \tag{60}$$

where $u=\tilde{u}-\tilde{\tilde{u}}$ and $g_k[\tilde{u},\tilde{\tilde{u}}]=g_k[\tilde{u}]-g_k[\tilde{\tilde{u}}]$. To estimate the integrals in (59), for each member h of the right side of (60), we distinguish seven cases (i)-(vii). Lemma 4 applies in all of them except (iv), when Lemma 3 does. In (ii)-(v) and (vii), the magnitude of the integrand h is estimated, before using Lemmas 3 and 4.

(i) $h=(p_k-q_k)u^{(n-k-1)}$, $p(t)=(p_k(t)-q_k(t))\,e^{ct}$, $\psi_1(t)=e^{ct}$, $\psi_2(t)=\exp[-(\operatorname{Re}\lambda_j+c)t]\varphi(t)$,

$$K(t)=\exp[-(\lambda_j+c)t]u^{(n-k-1)},\quad K_0=\|u\|(1+|\lambda_j+c|),\quad 2\leqq k\leqq n-1,\quad 1\leqq j\leqq L.$$

(ii) $h=q_k u^{(n-k-1)}$, $|h|\leqq\|u\|\varphi|q_k|$, $p(t)=|q_k(t)|\,e^{ct}\varphi(t)$, $\psi_1(t)=e^{ct}\varphi(t)$, $\psi_2(t)=\exp[-(\operatorname{Re}\lambda_j+c)t]$, $K=\|u\|\psi_2$, $K_0=\|u\|\max(1,|\operatorname{Re}\lambda_j+c|)$, $2\leqq k\leqq n=1$, $1\leqq j\leqq L$.

(iii) $h=p_1u^{(n-2)}$, $|h|\leqq\|u\|\varphi|p_1|$, $p(t)=|p_1(t)|\,e^{ct}\varphi(t)$, ψ_1, ψ_2, K and K_0 as in (ii), $1\leqq j\leqq L$.

(iv) $h=g_{n-k}[\tilde{u},\tilde{\tilde{u}}]$, $|h|\leqq C\|u\|\varphi^2$, $\sigma=C\|u\|\varphi$,

$$\psi(t)=\exp[-\operatorname{Re}\lambda_j t]\varphi(t),\quad 0\leqq k\leqq n-1,\quad 1\leqq j\leqq L.$$

(v) $h=p_1g_{n-1}[\tilde{u},\tilde{\tilde{u}}]$, $|h|\leqq C\|u\|\varphi^2|p_1|$, p as in (iii), ψ_1, ψ_2, K and K_0 as in (ii), $1\leqq j\leqq L$.

(vi) $h=(p_k-q_k)g_{n-k}[\tilde{u},\tilde{\tilde{u}}]$, p, ψ_1 and ψ_2 as in (ii), $K(t)=\exp[-(\lambda_j+c)t]\times g_{n-k}[\tilde{u},\tilde{\tilde{u}}]$, $K_0=C\|u\|(1+|\lambda_j+c|)$, $2\leqq k\leqq n-1$, $1\leqq j\leqq L$.

(vii) $h=q_kg_{n-k}[\tilde{u},\tilde{\tilde{u}}]$, $|h|\leqq C\|u\|\varphi^2|q_k|$, p, ψ_1, ψ_2, K and K_0 as in (ii), $2\leqq k\leqq n-1$, $1\leqq j\leqq L$.

Finally note that the existence of $\mathscr{L}[\tilde{u}]$ follows from the preceding with $\tilde{\tilde{u}}=0$, because $\mathscr{L}[\tilde{\tilde{u}}]=0$. This completes the proof of Lemma 5.

Now it is easy to finish the proof of Theorem 1. From (47), (53) and (54) with $\tilde{u} = u$ and $\tilde{\tilde{u}} = 0$, we find that

$$|D^i \mathcal{T}[u](t)| \leqq 2\sigma(t_0, t)\varphi(t) \qquad (0 \leqq i \leqq n-2) \tag{61}$$

for any u in $\mathcal{S}(t_0)$. Moreover, if $\tilde{u}$, $\tilde{\tilde{u}}$ are in $\mathcal{S}(t_0)$, then

$$|D^i \mathcal{T}[\tilde{u}](t) - D^i \mathcal{T}[\tilde{\tilde{u}}](t)| \leqq \|\tilde{u} - \tilde{\tilde{u}}\| \sigma(t_0, t)\varphi(t) \qquad (0 \leqq i \leqq n-2). \tag{62}$$

Since $\sigma(t_0, t)$ is nonincreasing in t, (48), (49), (61) and (62) imply that $\|\mathcal{T}[u]\| \leqq 2\sigma(t_0, t_0)$ and $\|\mathcal{T}[\tilde{u}] - \mathcal{T}[\tilde{\tilde{u}}]\| \leqq \sigma(t_0, t_0)\|\tilde{u} - \tilde{\tilde{u}}\|$. Therefore, $\mathcal{T}$ is a contraction mapping of $\mathcal{S}(t_0)$ into itself if t_0 is so large that $\sigma(t_0, t_0) \leqq 1/2$, which we now assume (recall that $\sigma(t_0, t_0) \to 0$ as $t_0 \to \infty$). Consequently, $\mathcal{T}$ has a fixed point (function) u_0 which satisfies $u_0 = \mathcal{T}u_0$. Setting $u = u_0$ in (61), we conclude that

$$D^k u_0(t) = o(\varphi(t)) \qquad (0 \leqq K \leqq n-2), \tag{63}$$

because $\sigma(t_0, t) \to 0$ as $t \to \infty$. Moreover, (38) and (51) with $u = u_0$ and $\tilde{\tilde{u}} = 0$ imply that

$$g_k[u_0](t) = o(\varphi(t)) \qquad (1 \leqq k \leqq n). \tag{64}$$

Since u_0 is a solution of (46), the function

$$x_0(t) = \exp\left[\int_{t_0}^{t} u_0(s)\, ds\right]$$

satisfies (1) on $[t_0, \infty)$. From (40) with $x = x_0$ and $u = u_0$, (63) and (64), we conclude that x_0 is as in (6) with $\lambda_0 = 0$. This completes the proof of Theorem 1 in the case when (38) holds.

Our proof requires (38), because of part (iv) in the proof of Lemma 5, where $\sigma(t) = C\|u\|\varphi(t) = o(1)$ is assumed. Moreover, (38) is necessary for (51) to imply (64). From the statement of Theorem 1, it is clear that if (38) does not hold, then it may as well be assumed that $\varphi = 1$. This case was considered in [6], proivded that the integrals $\int^\infty |p_1(t)|\, dt$ and $\int^\infty p_k(t)\, dt$ $(2 \leqq k \leqq n)$ converge and that (4) has n roots with distinct real parts. Fortunately, it presents no difficulty to adapt the proof of [6] to that of Theorem 1 with $\varphi = 1$, and we omit it here.

4. Corollaries of Theorem 1. First we give two immediate consequences of Theorem 1.

COROLLARY 1. *Theorem* 1 *remains valid if* (11)-(13) *are replaced by*

$$p_k(t) = o(1) \quad (1 \leqq k \leqq n), \qquad f(t) = o(\varphi(t)). \tag{65}$$

Proof. Obviously, (65) implies (11)-(13).

COROLLARY 2. *Theorem* 1 *remains valid if* (11)-(13) *are replaced by the assumption that the integrals* $\int^\infty |p_1(t)|\varphi(t)\, dt$, $\int^\infty p_k(t)\, dt$ $(2 \leqq k \leqq n)$ *and* $\int^\infty f(t)\, dt$ *converge so that*

$$\int_t^\infty |p_1(s)|\varphi(s)\, ds = o(\varphi(t)), \qquad \int_t^\infty f(s)\, ds = o(\varphi(t)).$$

Proof. From Lemma 1 with $\gamma = 0$, the assumption in Corollary 2 implies (11)-(13).

Corollary 2 with $\varphi(t) = e^{-\rho t}t^{-q}$ (ρ and q are nonnegative constants) and the conclusion of Remark 2 in § 1 cover a part of the author's results [5] and [6]. Equation (1) was considered there under assumptions including the convergence of the integrals $\int^\infty p_k(t)\, e^{\rho t}t^q\, dt$, which is more restrictive than

$$\int_t^\infty p_k(s)\, ds = o(e^{-\rho t}t^{-q}).$$

Very recently, Trench [8] has given conditions like those of Corollary 2 which imply the behavior (18) in the difficult case when $\operatorname{Re}(\lambda_0-\lambda_j)=m_\varphi$, for some some roots λ_j of (4). (All these λ_j are assumed to be simple.) However, Trench's assumptions include either that

$$\int_t^\infty \varphi^2(s)\,ds = O(\varphi(t)),$$

or some additional smallness restrictions on p_k.

The last consequence of Theorem 1 concerns the equation

$$x^{(n)}+p_1(t)x^{(n-1)}+\cdots+p_n(t)x=0 \qquad (t\geqq 0,\ n\geqq 2) \tag{66}$$

with continuous coefficients p_k, regarded as a perturbation of $x^{(n)}=0$. Trench [7] has considered a solution x_0 of (66) satisfying

$$x_0^{(k)}(t)=[Q(t)t^d]^{(k)}+o(t^{d-k}\psi(t)) \qquad (0\leqq k\leqq n-1),$$

where ψ is positive and nonincreasing on $[T,\infty)$, d is an integer $(0\leqq d\leqq n-1)$ and Q is a polynomial of degree $\leqq n-d-1$. We restrict our attention to the case $Q=1$. If we introduce the notation

$$\begin{aligned} Q_k(\lambda)&=\prod_{j=0}^{k-1}(\lambda-j), \qquad Q_0=1,\\ g(t)&=\sum_{k=1}^{n} Q_{n-k}(d)p_k(t)t^{d-n+k}, \end{aligned} \tag{67}$$

then Trench's result may be stated as follows.

THEOREM 2. *Suppose ψ is continuous, positive and nonincreasing on $[T,\infty)$, d is a fixed integer $(0\leqq d\leqq n-1)$ and also, if $d\neq 0$, suppose $t^\gamma\psi(t)$ is nondecreasing on $[T,\infty)$, for some $\gamma<1$. Assume also that the integrals $\int^\infty p_k(t)\,dt$ $(2\leqq k\leqq n)$ and $\int^\infty t^{n-d-1}g(t)\,dt$ converge and satisfy*

$$-\int_t^\infty p_k(s)\,ds = 0(t^{1-k}) \qquad (2\leqq k\leqq n), \tag{68}$$

and

$$\int_t^\infty s^{n-d-1}g(s)\,ds = o(\psi(t)). \tag{69}$$

Finally, suppose that

$$\int^\infty |p_1(t)|\,dt<\infty, \tag{70}$$

and

$$\int_t^\infty s^{k-2}\psi(s)\left|\int_s^\infty p_k(r)\,dr\right| ds = o(\psi(t)) \qquad (2\leqq k\leqq n). \tag{71}$$

Then (66) has a solution x_0 satisfying

$$x_0^{(k)}(t)=[Q_k(d)+o(\psi(t))]t^{d-k} \quad (0\leqq k\leqq n-1).$$

Our following result implies that even (71) need not hold. Note first that, by Lemma 2, (68) is equivalent to

$$\sup_{s\geqq t}(1+\log(s/t))^{-1}\left|\int_t^s r^{k-1}p_k(r)\,dr\right|=o(1)\qquad(2\leqq k\leqq n),\tag{72}$$

while

$$\sup_{s\geqq t}(1+\log(s/t))^{-1}\left|\int_t^s r^{n-d-1}g(r)\,dr\right|=o(\psi(t)),\tag{73}$$

and

$$\sup_{s\geqq t}(1+\log(s/t))^{-1}\int_t^s |p_1(r)|\psi(r)\,dr=o(\psi(t))\tag{74}$$

are weaker than (69) and (70), respectively.

COROLLARY 3. *Let fixed ψ and d be as in the first sentence of Theorem* 2. *If* (72)–(74) *hold, then* (66) *has a solution x_0 satisfying*

$$\frac{x_0^{(k)}(t)}{x_0(t)}=[Q_k(d)+o(\psi(t))]t^{-k}\qquad(1\leqq k\leqq n-1).\tag{75}$$

Proof. We introduce the new variables y, τ by

$$\tau=\log t,\qquad x(t)=y(\tau).\tag{76}$$

Then

$$x^{(k)}(t)=e^{-k\tau}Q_k(D)y(\tau)\qquad\left(D=\frac{d}{d\tau},\,0\leqq k\leqq n\right)\tag{77}$$

and therefore, (66) is transformed into

$$Q_n(D)y+\sum_{k=1}^{n}\tilde{p}_k(\tau)D^{n-k}y=0,\tag{78}$$

where

$$\tilde{p}_k(\tau)=\frac{1}{(n-k)!}\sum_{j=1}^{k}Q_{n-j}^{(n-k)}(0)p_j(e^\tau)\,e^{j\tau}\qquad(1\leqq k\leqq n).\tag{79}$$

We now verify that (78) satisfies the assumptions of Theorem 1 with t replaced by τ, $\varphi(\tau)=\psi(e^\tau)$ and $\lambda_0=d$. Indeed, (4) now becomes $Q_n(\lambda)=0$, and hence has n distinct roots $\lambda_j=j-1$, $1\leqq j\leqq n$. If $d>0$, then $M_\varphi\leqq\gamma<1$, because $\varphi(\tau)\,e^{\gamma\tau}(=\psi(t)t^\gamma)$ is nondecreasing on $[\log\max(1,T),\infty]$. Thus (9) holds with $\lambda_0=d$, for any $\lambda_j\neq d$. Routine manipulations with (67) and (79) show that f in (14) now becomes $f(\tau)=g(e^\tau)\exp[(n-d)\tau]$. From this and (79), substituting $r=\log u$ yields

$$\int_\tau^s|\tilde{p}_1(r)|\varphi(r)\,dr=\int_{e^\tau}^{e^s}|p_1(u)|\psi(u)\,du,$$

$$\int_\tau^s f(r)\,dr=\int_{e^\tau}^{e^s}u^{n-d-1}g(u)\,du$$

and

$$\int_\tau^s(\tilde{p}_k(r)-c_1^k\tilde{p}_1(r))dr=\frac{1}{(n-k)!}\sum_{j=2}^{k}c_j^k\int_{e^\tau}^{e^s}u^{j-1}p_j(u)\,du\qquad(2\leqq k\leqq n),$$

where $c_j^k = Q_{n-j}^{(n-k)}(0)$. Consequently, (72)-(74) imply (11), (13), (15) and (16) with $p_k = \tilde{p}_k$ and $q_k = c_1^k \tilde{p}_1 (2 \leqq k \leqq n)$. Theorem 1 implies that (78) has a nontrivial solution y_0 satisfying

$$y_0^{(k)}(\tau) = [d^k + o(\varphi(\tau))] y_0(\tau) \qquad (1 \leqq k \leqq n-1)$$

as $\tau \to \infty$. Then (77) shows that the solution $x_0(t) = y_0(\log t)$ of (66) satisfies (75). This completes the proof of Corollary 3.

REFERENCES

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

[2] ———, *Unrestricted solution fields of almost separable differential equations*, Trans. Amer. Math. Soc., 63 (1948), pp. 560-580.

[3] P. HARTMAN AND A. WINTNER, *Asymptotic integrations of linear differential equations*, Amer. J. Math., 77 (1955), pp. 45-87.

[4] O. PERRON, *Ueber lineare Differentialgleichungen, bei denen die unabhangig Variable reele ist*, J. Reine Angew. Math., 142 (1913), pp. 254-270.

[5] J. ŠIMŠA, *Asymptotic integration of perturbed linear differential equations under conditions involving ordinary integral convergence*, this Journal, 15 (1984), pp. 116-123.

[6] ———, *The condition of ordinary integral convergence in the asymptotic theory of linear differential equations with almost constant coefficients*, this Journal, 16 (1985), pp. 757-769.

[7] W. F. TRENCH, *Asymptotic integration of linear differential equations subject to mild integral conditions*, this Journal, 15 (1984), pp. 932-942.

[8] ———, *Linear perturbations of a constant coefficient differential equation subject to mild integral smallness conditions*, Czech. Math. J., 36 (1986), pp. 623-633.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

015

POLYNOMIALS RELATED TO EXPANSIONS OF CERTAIN RATIONAL FUNCTIONS IN TWO VARIABLES*

KARL DILCHER†

Abstract. A difference equation corresponding to a certain partial differential equation leads to a "Pascal type" triangle. The entries of a row of this triangle can be regarded as coefficients of a polynomial; the sequence of these polynomials is studied, together with its generating function and related polynomials. The entries of a more general class of number triangles are explicitly determined, as well as asymptotic expressions for the columns of the triangles. Chebyshev and Gegenhauer polynomials, as well as hypergeometric functions are used in the proofs.

Key words. sequences of polynomials, recursions, zeros of polynomials, Chebyshev polynomials, ultraspherical polynomials, hypergeometric functions, asymptotics, Darboux's method

AMS(MOS) subject classifications. 11B37, 33A30, 33A50, 41A60

1. Introduction.

Let $u = u(x, t)$ be a function in two variables, and consider the (hyperbolic) partial differential equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2} + \frac{\partial u}{\partial t}.$$

If we change this into a difference equation, we get

$$2u(x, t+1) = u(x-1, t) + u(x, t) + u(x+1, t) - u(x, t-1).$$

This suggests the "Pascal type" triangle (after normalizing)

(1.1)

$$\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & 1 & 1 & & & & \\
 & & & 1 & 2 & 1 & 2 & 1 & & & \\
 & & 1 & 3 & 2 & 3 & 2 & 3 & 1 & & \\
 & 1 & 4 & 4 & 4 & 5 & 4 & 4 & 4 & 1 & \\
1 & 5 & 7 & 6 & 9 & 7 & 9 & 6 & 7 & 5 & 1 \\
 & & & & & \vdots & & & & &
\end{array}$$

where each element in the nth row is the sum of the three closest elements in the $(n-1)$th row, minus twice the closest element in the $(n-2)$th row.

Now we expand

$$G(z, t) := \frac{t}{1 - t(1+z+z^2) + 2z^2t^2} = \sum_{n=1}^{\infty} f_n(z) t^n;$$

it is clear that the $f_n(z)$ are polynomials of degree $2n$, and their coefficients are the rows of the triangle (1.1).

More generally, let $\nu > \frac{1}{2}$ and λ be real parameters. We expand

$$G^{\lambda,\nu}(z, t) := (1 - (1+z+z^2)t + \lambda z^2 t^2)^{-\nu} = \sum_{n=0}^{\infty} f_n^{\lambda,\nu}(z) t^n.$$

* Received by the editors January 13, 1986; accepted for publication February 9, 1987.

† Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5. The work of this author was supported by a Killam Postdoctoral Fellowship.

If we compare this with the generating function

$$(1-2zt+t^2)^{-\nu}=\sum_{n=0}^{\infty} C_n^{\nu}(z)t^n$$

for the ultraspherical (Gegenhauer) polynomials $C_n^{\nu}(z)$, we find

$$f_n^{\lambda,\nu}(z)=\lambda^{n/2}z^n C_n^{\nu}\left(\frac{1+z+z^2}{2\sqrt{\lambda}z}\right). \tag{1.2}$$

Using the recurrence relation for the ultraspherical polynomials (see, e.g., [1, p. 782]), we get $f_0^{\lambda,\nu}(z)=1$, $f_1^{\lambda,\nu}(z)=\nu(1+z+z^2)$, and

$$f_n^{\lambda,\nu}(z)=\left(1+\frac{\nu-1}{n}\right)(1+z+z^2)f_{n-1}^{\lambda,\nu}(z)-\left(1+2\frac{\nu-1}{n}\right)\lambda z^2 f_{n-2}^{\lambda,\nu}(z). \tag{1.3}$$

The polynomials $f_n^{\lambda,\nu}(z)$ are self-inverse, i.e., $f_n^{\lambda,\nu}(z)=z^{2n}f_n^{\lambda,\nu}(1/z)$. If we denote

$$f_n^{\lambda,\nu}(z)=C_{n,n}^{\lambda,\nu}+C_{n,n-1}^{\lambda,\nu}z+\cdots+C_{n,0}^{\lambda,\nu}z^n+C_{n,1}^{\lambda,\nu}z^{n+1}+\cdots+C_{n,n}^{\lambda,\nu}z^{2n}, \tag{1.4}$$

we get the triangle

$$\begin{array}{c} C_{0,0}^{\lambda,\nu} \\ C_{1,1}^{\lambda,\nu}\,C_{1,0}^{\lambda,\nu}\,C_{1,1}^{\lambda,\nu} \\ C_{2,2}^{\lambda,\nu}\,C_{2,1}^{\lambda,\nu}\,C_{2,0}^{\lambda,\nu}\,C_{2,1}^{\lambda,\nu}\,C_{2,2}^{\lambda,\nu} \\ \vdots \end{array} \tag{1.5}$$

where

$$C_{n,k}^{\lambda,\nu}=\left(1+\frac{\nu-1}{n}\right)(C_{n-1,k-1}^{\lambda,\nu}+C_{n-1,k}^{\lambda,\nu}+C_{n-1,k+1}^{\lambda,\nu})-\left(1+2\frac{\nu-1}{n}\right)\lambda C_{n-2,k}^{\lambda,\nu}, \tag{1.6}$$

with $C_{n,k}^{\lambda,\nu}=C_{n,-k}^{\lambda,\nu}$. For $\lambda=2$ and $\nu=1$, the triangle (1.5) has the form (1.1).

The main purpose of this paper is to study the coefficients $C_{n,k}^{\lambda,\nu}$. We derive the following explicit and asymptotic expressions.

THEOREM 1.

$$C_{n,k}^{\lambda,\nu}=\frac{1}{\Gamma(\nu)}\sum_{s=0}^{[(n-k)/2]}(-\lambda)^s\frac{\Gamma(\nu+n-s)}{s!(n-2s)!}\sum_{j=0}^{[(n-k-2s)/2]}\binom{2j+k}{j}\binom{n-2s}{2j+k}.$$

Here $[x]$ denotes, as usual, the greatest integer function.

THEOREM 2. *For fixed real $\nu>\frac{1}{2}$ and λ, and integer $k\geqq 0$, we have asymptotically as $n\to\infty$,*

(a) *if $\lambda<\frac{9}{4}$,*

$$C_{n,k}^{\lambda,\nu}\sim\frac{1}{2n\Gamma(\nu)\sqrt{\pi}}\left(\frac{n}{\sqrt{9-4\lambda}}\right)^{\nu-1/2}\left(\frac{3+\sqrt{9-4\lambda}}{2}\right)^{n+\nu};$$

(b) *if $\lambda>\frac{9}{4}$,*

$$C_{n,k}^{\lambda,\nu}\sim\frac{(\sqrt{\lambda})^{n+\nu}n^{\nu-3/2}}{\Gamma(\nu)\sqrt{\pi}}\left\{(\sqrt{4\lambda-1})^{1/2-\nu}\cos\left[(\alpha+\pi)n+\nu\alpha+\left(k+\frac{1}{4}-\frac{3\nu}{2}\right)\pi\right]\right.$$

$$\left.+(\sqrt{4\lambda-9})^{1/2-\nu}\cos\left[\beta n+\nu\beta+\left(\frac{1}{4}-\frac{\nu}{2}\right)\pi\right]\right\},$$

where

$$\alpha = \mathrm{Cos}^{-1}(1/2\sqrt{\lambda}), \qquad \beta = \mathrm{Cos}^{-1}(3/2\sqrt{\lambda});$$

(c) *if* $\lambda = \frac{9}{4}$,

$$C_{n,k}^{\lambda,\nu} \sim \frac{(3/2)^n}{\Gamma(\nu)}\left\{\frac{\sqrt{2/3}}{\Gamma(\nu)}\left(\frac{n}{2}\right)^{2\nu-2} + \frac{(\sqrt{8})^{1/2-\nu}}{\sqrt{\pi}}\left(\frac{3}{2}\right)^{\nu} n^{\nu-3/2}\cos\left[(\alpha+\pi)n + \nu\alpha + \left(k+\frac{1}{4}-\frac{3\nu}{2}\right)\pi\right]\right\},$$

where

$$\alpha = \mathrm{Cos}^{-1}(\tfrac{1}{3}).$$

Next we fix $\nu = 1$ and expand $G^{\lambda,1}(z, t)$ according to powers of z. If we set $t = 1 - \alpha^{-1}$, we get

$$G^{\lambda,1}(z, t) = \frac{\alpha}{1-(\alpha-1)z+\{(\lambda-1)\alpha+(1-2\lambda)+\lambda\alpha^{-1}\}z^2} = \alpha \sum_{n=0}^{\infty} g_n^{\lambda}(\alpha) z^n.$$

We have $g_0^{\lambda}(\alpha) = 1$, $g_1^{\lambda}(\alpha) = \alpha - 1$, and

$$g_{n+1}^{\lambda}(\alpha) = (\alpha-1)\{g_n^{\lambda}(\alpha) + (\lambda\alpha^{-1}+1-\lambda)g_{n-1}^{\lambda}(\alpha)\}. \tag{1.7}$$

In § 2 we find explicit expressions for the zeros of the $f_n^{\lambda,1}(z)$ and the $g_n^{\lambda,1}(\alpha)$ for all values of λ. In §§ 3-5, Theorems 1 and 2 are proved, and § 6 contains some further remarks and generalizations.

2. The zeros. The Chebyshev polynomials of the second kind $U_n(z)$ can be defined by the recursion $U_0(z) = 1$, $U_1(z) = 2z$, and

$$U_{n+1}(z) = 2zU_n(z) - U_{n-1}(z). \tag{2.1}$$

By taking $z = p(x)/2\sqrt{q(x)}$ we get the following lemma.

LEMMA 1. *Let* $p(x)$ *and* $q(x)$ *be arbitrary functions, and define the sequence* $V_n(x)$ *recursively by* $V_0(x) = 1$, $V_1(x) = p(x)$, *and*

$$V_{n+1}(x) = p(x)V_n(x) - q(x)V_{n-1}(x).$$

Then, if $q(x) \neq 0$,

$$V_n(x) = q(x)^{n/2} U_n(p(x)/2\sqrt{q(x)}).$$

To find the zeros of $g_n^{\lambda}(\alpha)$, we take $p(\alpha) := \alpha - 1$ and $q(\alpha) := -(\alpha-1)(\lambda\alpha^{-1}+1-\lambda)$. With (1.7) and Lemma 1 we find that

$$g_n^{\lambda}(\alpha) = \{(\alpha-1)(\lambda\alpha^{-1}+1-\lambda)\}^{n/2} U_n\left(\frac{i}{2}\left(\frac{\alpha-1}{\lambda\alpha^{-1}+1-\lambda}\right)^{1/2}\right).$$

Hence $g_n^{\lambda}(\alpha)$ has zeros when

$$\frac{\alpha-1}{\lambda\alpha^{-1}+1-\lambda} = -4\cos^2\frac{k\pi}{n+1},$$

i.e., the zeros are given by

$$\alpha = \frac{1}{2} - 2(1-\lambda)\cos^2\frac{k\pi}{n+1} \pm \left\{4(1-\lambda^2)\cos^4\left(\frac{k\pi}{n+1}\right) - 2(1+\lambda)\cos^2\left(\frac{k\pi}{n+1}\right) + \frac{1}{4}\right\}^{1/2}$$

$(k=1,2,\cdots,n)$. It is easy to see that these zeros are real unless

$$\frac{1+\lambda-2\sqrt{\lambda}}{4(\lambda-1)^2}<\sin^2\left(\frac{k\pi}{n+1}\right)<\frac{1+\lambda+2\sqrt{\lambda}}{4(\lambda-1)^2},$$

in which case they lie on the circle

$$y^2+\left(x-\frac{\lambda}{\lambda-1}\right)^2=\left(\frac{\sqrt{\lambda}}{\lambda-1}\right)^2 \qquad (\alpha=x+iy).$$

To find the zeros of $f_n^{\lambda,1}(z)$, we use the facts that $C_n^1(z)=U_n(z)$, and that the zeros of $U_n(z)$ are given by $\cos(k\pi/(n+1))$, $k=1,2,\cdots,n$. Hence with (1.2) we find that the $2n$ zeros of $f_n^{\lambda,1}(z)$ are

$$z=-\sqrt{\lambda}\cos\frac{k\pi}{n+1}-\frac{1}{2}\pm\left(\lambda\cos^2\frac{k\pi}{n+1}+\sqrt{\lambda}\cos\frac{k\pi}{n+1}-\frac{3}{4}\right)^{1/2}$$

for $k=1,2,\cdots,n$. We note that these zeros are real except when

$$\frac{-3}{2\sqrt{\lambda}}<\cos\frac{k\pi}{n+1}<\frac{1}{2\sqrt{\lambda}},$$

in which case they lie on the unit circle.

3. Proof of Theorem 1. Using the well-known explicit expression for the ultraspherical polynomials (see, e.g., [1, p. 775]) and (1.2), we get

$$f_n^{\lambda,\nu}(z)=\frac{1}{\Gamma(\nu)}\sum_{s=0}^{[n/2]}(-\lambda)^s\frac{\Gamma(\nu+n-s)}{s!(n-2s)!}z^{2s}(1+z+z^2)^{n-2s}. \tag{3.1}$$

If r is a positive integer, the binomial theorem, applied twice, gives

$$\begin{aligned}(1+z+z^2)^r&=\sum_{j=0}^{r}\sum_{i=0}^{j}\binom{r}{j}\binom{j}{i}z^{2j-i}\\&=\sum_{m=0}^{2r}z^m\sum_{j=0}^{[m/2]}\binom{r}{m-j}\binom{m-j}{m-2j},\end{aligned}$$

and with (3.1) we obtain

$$\begin{aligned}f_n^{\lambda,\nu}(z)&=\frac{1}{\Gamma(\nu)}\sum_{s=0}^{[n/2]}(-\lambda)^s\frac{\Gamma(\nu+n-s)}{s!(n-2s)!}z^{2s}\sum_{m=0}^{2n-4s}z^m\sum_{j=0}^{[m/2]}\binom{n-2s}{m-j}\binom{m-j}{m-2j}\\&=\frac{1}{\Gamma(\nu)}\sum_{k=-n}^{n}z^{n-k}\sum_{s=0}^{[(n-k)/2]}(-\lambda)^s\frac{\Gamma(\nu+n-s)}{s!(n-2s)!}\sum_{j=0}^{[(n-k-2s)/2]}\binom{n-2s}{n-k-j-2s}\\&\quad\cdot\binom{n-k-j-2s}{n-k-2j-2s}.\end{aligned}$$

The theorem now follows if we compare the last equation with (1.4) and note that the product of the two binomial coefficients in the last line is equal to that in Theorem 1.

4. Lemmas. We can rewrite Theorem 1 in the form

$$C_{n,k}^{\lambda,\nu}=\frac{1}{\Gamma(\nu)}\sum_{s=0}^{[(n-k)/2]}(-\lambda)^s\binom{n-k-s}{s}\frac{\Gamma(\nu+n-s)}{k!(n-k-s)!}B_k^{(n-k-2s)}, \tag{4.1}$$

where

$$B_k^{(m)} := \sum_{j=0}^{[m/2]} \binom{2j}{j}\binom{m}{2j}\Big/\binom{k+j}{j}.$$

LEMMA 2.

$$B_k^{(m)} = (-i\sqrt{3})^m \frac{m!(2k)!}{(m+2k)!} C_m^{k+1/2}\left(\frac{i}{\sqrt{3}}\right).$$

Proof. We have

$$\begin{aligned} B_k^{(m)} &= \sum_{j=0}^{[m/2]} \frac{(2j)!\,m!\,j!\,k!}{j!j!(m-2j)!(2j)!(k+j)!} \\ &= \sum_{j=0}^{[m/2]} \frac{(-m)_{2j}}{(k+1)_j j!} = \sum_{j=0}^{[m/2]} \frac{(-m/2)_j((1-m)/2)_j}{(k+1)_j} \frac{2^{2j}}{j!} \\ &= F\left(-\frac{m}{2}, \frac{1-m}{2}; k+1; 4\right), \end{aligned}$$

where $(a)_j$ is the Pochhammer symbol $(a)_j = a(a+1)\cdot\cdots\cdot(a+j-1)$ and $(a)_0 = 1$ and $F(a, b; c; x) = {}_2F_1(a, b; c; x)$ is the Gauss hypergeometric series (see, e.g., [1, p. 556]). The ultraspherical polynomials can be expressed as

$$C_m^{\nu}(x) = \frac{(2\nu)_m}{m!} x^m F\left(-\frac{m}{2}, \frac{1-m}{2}; \nu+\frac{1}{2}; 1-x^{-2}\right);$$

this gives the lemma, with $\nu = k+\frac{1}{2}$, $x = i/\sqrt{3}$.

If we combine (4.1) with Lemma 2, we get

LEMMA 3.

$$C_{n,k}^{\lambda,\nu} = \frac{(2k)!}{\Gamma(\nu)k!}(-i\sqrt{3})^{n-k} \sum_{s=0}^{[(n-k)/2]} \left(\frac{\lambda}{3}\right)^s \frac{\Gamma(\nu+n-s)}{s!(n+k-2s)!} C_{n-k-2s}^{k+1/2}\left(\frac{i}{\sqrt{3}}\right).$$

5. Proof of Theorem 2. First we determine the generating functions for the $C_{n,k}^{\lambda,\nu}$, where k, λ, and ν are fixed. To simplify notation, we write $C_n := C_{n,k}^{\lambda,\nu}$. We denote

$$d_n := \frac{\Gamma(\nu)k!}{\Gamma(k+\nu)}(\sqrt{\lambda})^{k-n} C_n \tag{5.1}$$

and

$$t_1 := \frac{-1}{2\sqrt{\lambda}}(1-\sqrt{1-4\lambda}), \qquad t_2 := \frac{-1}{2\sqrt{\lambda}}(1+\sqrt{1-4\lambda}),$$

$$t_3 := \frac{1}{2\sqrt{\lambda}}(3+\sqrt{9-4\lambda}), \qquad t_4 := \frac{1}{2\sqrt{\lambda}}(3-\sqrt{9-4\lambda});$$

note that $t_1 t_2 = 1 = t_3 t_4$.

LEMMA 4. *For real $\nu > 1/2$ and λ, and for $k = 0, 1, \cdots$, we have*

$$F_k(t) := t^k[(1+tt_1)(1+tt_2)]^{\nu-k-1}[(1-tt_1)(1-tt_2)(1-tt_3)(1-tt_4)]^{1/2-\nu}$$
$$\cdot F\left(\frac{k-\nu+2}{2}, \frac{k-\nu+1}{2}; k+1; \frac{4t^2/\lambda}{(t^2-(t/\sqrt{\lambda})+1)^2}\right) = \sum_{n=k}^{\infty} d_n t^n.$$

Proof. We use Lemma 3, change the order of summation, and apply the binomial theorem

$$\begin{aligned}
\sum_{n=k}^{\infty}\left(\frac{i}{\sqrt{3}}\right)^{n-k}\frac{\Gamma(\nu)k!}{(2k)!}C_n t^n &= \sum_{n=k}^{\infty}\left\{\sum_{s=0}^{[(n-k)/2]}\left(\frac{\lambda}{3}\right)^s\frac{\Gamma(\nu+n-s)}{s!(n+k-2s)!}C_{n-k-2s}^{k+1/2}\left(\frac{i}{\sqrt{3}}\right)\right\}t^n \\
&= \sum_{m=k}^{\infty}C_{m-k}^{k+1/2}\left(\frac{i}{\sqrt{3}}\right)t^m\sum_{s=0}^{\infty}\frac{\Gamma(\nu+m+s)}{s!(m+k)!}\left(\frac{\lambda}{3}t^2\right)^s \\
&= \sum_{m=k}^{\infty}C_{m-k}^{k+1/2}\left(\frac{i}{\sqrt{3}}\right)t^m\frac{\Gamma(\nu+m)}{(m+k)!}\sum_{s=0}^{\infty}\frac{\Gamma(\nu+m+s)}{s!\Gamma(\nu+m)}\left(\frac{\lambda}{3}t^2\right)^s \\
&= \sum_{m=k}^{\infty}C_{m-k}^{k+1/2}\left(\frac{i}{\sqrt{3}}\right)\frac{\Gamma(\nu+m)}{(m+k)!}t^m\left(1-\frac{\lambda}{3}t^2\right)^{-m-\nu} \\
&= \frac{t^k}{(1-(\lambda/3)t^2)^{k+\nu}}\frac{\Gamma(k+\nu)}{(2k)!}\sum_{m=0}^{\infty}C_m^{k+1/2}\left(\frac{i}{\sqrt{3}}\right) \\
&\quad\cdot\frac{(k+\nu)_m}{(2k+1)_m}\left(\frac{t}{1-(\lambda/3)t^2}\right)^m,
\end{aligned}$$

where we have used $(x)_n=\Gamma(x+n)/\Gamma(x)$. After changing the variable t to $-it\sqrt{3/\lambda}$, we get

$$\sum_{n=k}^{\infty}d_n t^n=\frac{t^k}{(1+t^2)^{k+\nu}}\sum_{m=0}^{\infty}C_m^{k+1/2}\left(\frac{i}{\sqrt{3}}\right)\frac{(k+\nu)_m}{(2k+1)_m}\left(\frac{-it\sqrt{3/\lambda}}{1+t^2}\right)^m. \tag{5.2}$$

Now we are going to use the following generating function for ultraspherical polynomials (see, e.g., [3, p. 279]),

$$\sum_{m=0}^{\infty}\frac{(\gamma)_m}{(2\alpha)_m}C_m^{\alpha}(x)z^m=(1-xz)^{-\gamma}F\left(\frac{\gamma}{2},\frac{\gamma+1}{2};\alpha+\frac{1}{2};\frac{z^2(x^2-1)}{(1-xz)^2}\right). \tag{5.3}$$

With $\gamma=k+\nu$, $\alpha=k+\frac{1}{2}$, $x=i\sqrt{3}$, and $z=-it\sqrt{3/\lambda}/(1+t^2)$, (5.2) becomes

$$\sum_{n=k}^{\infty}d_n t^n=\frac{t^k}{(t^2-(t/\sqrt{\lambda})+1)^{k+\nu}}F\left(\frac{k+\nu}{2},\frac{k+\nu+1}{2};k+1;y\right), \tag{5.4}$$

where

$$y=\frac{4t^2/\lambda}{(t^2-(t/\sqrt{\lambda})+1)^2}.$$

Using Euler's identity

$$F(a,b;c;y)=(1-y)^{c-a-b}F(c-a,c-b;c;y)$$

(for $|y|<1$; see [3, p. 60]), we get

$$F\left(\frac{k+\nu}{2},\frac{k+\nu+1}{2};k+1;y\right)=(1-y)^{1/2-\nu}F\left(\frac{k-\nu+2}{2},\frac{k-\nu+1}{2};k+1;y\right). \tag{5.5}$$

Now it is easy to verify that

$$1-y=[(1+tt_1)(1+tt_2)]^{-2}(1-tt_1)(1-tt_2)(1-tt_3)(1-tt_4);$$

the lemma now follows from (5.4) and (5.5).

Proof of Theorem 2. To find asymptotics for d_n, we apply Darboux's method (see, e.g., [2, p. 310]) on the generating function $F_k(t)$. Possible singularities of $F_k(t)$ are

at $t=t_j$, $j=1,\cdots,4$, and at $t=-t_1$, $t=-t_2$. To examine the behaviour of $F_k(t)$ in the neighbourhood of $-t_1$ and $-t_2$, we apply the identity (see, e.g., [1, p. 559])

$$F(a,b;c;z)=d_1(-z)^{-a}F\left(a,1-c+a;1-b+a;\frac{1}{z}\right)$$
$$+d_2(-z)^{-b}F\left(b,1-c+b;1-a+b;\frac{1}{z}\right),$$

where d_1 and d_2 are constants depending on a, b, c, to the right-hand side of (5.4). We find

$$F_k(t)=d_1\left(\frac{-\lambda}{4}\right)^{(k+\nu)/2}t^{-\nu}F\left(\frac{k+\nu}{2},\frac{-k+\nu}{2};\frac{1}{2};\frac{1}{y}\right)$$
$$+d_2\left(\frac{-\lambda}{4}\right)^{(k+\nu+1)/2}t^{-\nu-1}\left(t^2-\frac{t}{\sqrt{\lambda}}+1\right)F\left(\frac{k+\nu+1}{2},\frac{-k+\nu+1}{2};\frac{3}{2};\frac{1}{y}\right)$$

which shows that the singularities at $t=-t_1$ and $-t_2$ are removable.

Now we denote

$$F_k^{(j)}(t):=(1-tt_j)^{\nu-1/2}F_k(t)\qquad(j=1,\cdots,4)$$

and

$$F_k^{(5)}(t):=\{(1-tt_3)(1-tt_4)\}^{\nu-1/2}F_k(t).$$

Using the Gauss summation formula (see, e.g., [3, p. 49])

$$F\left(\frac{k-\nu+2}{2},\frac{k-\nu+1}{2};k+1;1\right)=\frac{\Gamma(k+1)\Gamma(\nu-1/2)}{\Gamma\left(\frac{k+\nu}{2}\right)\Gamma\left(\frac{k+\nu+1}{2}\right)}:=\Gamma,$$

which holds for $\nu>\frac{1}{2}$, we find

$$F_k^{(1)}(t_2)=\left(\frac{-1}{2}\right)^{k+\nu}(-\sqrt{1-4\lambda})^{1/2-\nu}(\sqrt{\lambda})^{k+\nu}\left(\frac{-1+\sqrt{1-4\lambda}}{2\sqrt{\lambda}}\right)^{\nu}\Gamma, \tag{5.6}$$

$$F_k^{(2)}(t_1)=\left(\frac{-1}{2}\right)^{k+\nu}(\sqrt{1-4\lambda})^{1/2-\nu}(\sqrt{\lambda})^{k+\nu}\left(\frac{-1-\sqrt{1-4\lambda}}{2\sqrt{\lambda}}\right)^{\nu}\Gamma, \tag{5.7}$$

$$F_k^{(3)}(t_4)=\left(\frac{1}{2}\right)^{k+\nu}(\sqrt{9-4\lambda})^{1/2-\nu}(\sqrt{\lambda})^{k+\nu}\left(\frac{3+\sqrt{9-4\lambda}}{2\sqrt{\lambda}}\right)^{\nu}\Gamma, \tag{5.8}$$

$$F_k^{(4)}(t_3)=\left(\frac{1}{2}\right)^{k+\nu}(-\sqrt{9-4\lambda})^{1/2-\nu}(\sqrt{\lambda})^{k+\nu}\left(\frac{3-\sqrt{9-4\lambda}}{2\sqrt{\lambda}}\right)^{\nu}\Gamma; \tag{5.9}$$

the arguments of these (in general multi-valued) expressions will be determined later.

We note that

(a) if $\lambda<\frac{9}{4}$, then $|t_4|<|t_j|$ for $j=1,2,3$;

(b) if $\lambda>\frac{9}{4}$, then $|t_j|=1$ $(j=1,\cdots,4)$ and no two t_j are equal;

(c) if $\lambda=\frac{9}{4}$, then $t_3=t_4=1$, $|t_1|=|t_2|=1$, $t_1\neq t_2$, $t_1\neq 1$, $t_2\neq 1$.

We prove Theorem 2 according to this distinction.

(a) Let $\lambda<\frac{9}{4}$. Then according to Darboux's method the coefficients in the MacLaurin expansion of

$$f(t):=F_k^{(3)}(t_4)(1-tt_3)^{1/2-\nu}$$

are asymptotics to the d_n, as $n\to\infty$. The binomial theorem gives

$$(1-tt_3)^{1/2-\nu}=\sum_{n=0}^{\infty}\frac{\Gamma(\nu-1/2+n)}{\Gamma(\nu-1/2)n!}\left(\frac{3+\sqrt{9-4\lambda}}{2\sqrt{\lambda}}\right)^{n}t^n, \tag{5.10}$$

and we obtain from (5.1), (5.8), and (5.10)

$$C_n \sim A2^{-k-\nu}(\sqrt{9-4\lambda})^{1/2-\nu}\left(\frac{3+\sqrt{9-4\lambda}}{2}\right)^{n+\nu}, \tag{5.11}$$

where all the gamma function and factorial terms from (5.1), (5.8) and (5.10) are collected in A. Using the duplication formula (see, e.g., [3, p. 24])

$$\frac{\Gamma(2z)}{\Gamma(z)\Gamma(z+\frac{1}{2})} = \frac{2^{2z-1}}{\sqrt{\pi}}, \tag{5.12}$$

we find

$$A = \frac{\Gamma(\nu-\frac{1}{2}+n)}{\Gamma(\nu)n!}\,\frac{2^{2k+\nu-1}}{\sqrt{\pi}}. \tag{5.13}$$

Stirling's formula now gives

$$\Gamma(\nu-\tfrac{1}{2}+n)/n! \sim n^{\nu-3/2} \qquad (\text{as } n\to\infty),$$

so finally we get with (5.11) and (5.13)

$$C_n \sim \frac{1}{\Gamma(\nu)\sqrt{\pi}}n^{\nu-3/2}(9-4\lambda)^{1/4-\nu/2}\left(\frac{3+\sqrt{9-4\lambda}}{2}\right)^{n+\nu},$$

which implies Theorem 2(a).

(b) Let $\lambda > \frac{9}{4}$. Asymptotics to the d_n (as $n\to\infty$) are given by the coefficients of the expansion of

$$g(t) := \sum_{j=1}^{4} F_k^{(j)}(t_j^{-1})(1-tt_j)^{1/2-\nu}.$$

We note that, in general, the values in (5.6)–(5.9) are not uniquely determined. However, the powers $(1-xz)^{-\gamma}$ and $(1-y)^{1/2-\nu}$ in (5.3), resp. (5.5) are to be taken with their principal values. With this in mind, we find that we have to take (5.6) and (5.7) with arguments

$$\varepsilon_1 := k\pi - \nu\alpha - \frac{\pi}{2}\left(\frac{1}{2}-\nu\right), \qquad \varepsilon_2 := k\pi + \nu\alpha + \frac{\pi}{2}\left(\frac{1}{2}-\nu\right),$$

respectively, where $\alpha := \arg\left((1+i\sqrt{4\lambda-1})/2\sqrt{\lambda}\right)$. Using the equivalent of (5.10) for t_1 and t_2, and with (5.1), (5.6), and (5.7), we find that the combined contribution from the first and second term of $g(t)$ is

$$\begin{aligned} &A2^{-k-\nu}(\sqrt{4\lambda-1})^{1/2-\nu}(\sqrt{\lambda})^{n+\nu}\left\{e^{i\varepsilon_1}\left(\frac{-1+i\sqrt{4\lambda-1}}{2\sqrt{\lambda}}\right)^n + e^{i\varepsilon_2}\left(\frac{-1-i\sqrt{4\lambda-1}}{2\sqrt{\lambda}}\right)^n\right\} \\ &\sim \frac{n^{\nu-3/2}}{\Gamma(\nu)\sqrt{\pi}}(\sqrt{4\lambda-1})^{1/2-\nu}(\sqrt{\lambda})^{n+\nu}\cos\left\{(\alpha+\pi)n+\nu\alpha+\left(k+\frac{1}{4}-\frac{3\nu}{2}\right)\pi\right\}. \end{aligned} \tag{5.14}$$

Similarly, we find that we have to take (5.8) and (5.9) with arguments

$$\varepsilon_3 := \nu\beta + \frac{\pi}{2}\left(\frac{1}{2}-\nu\right), \qquad \varepsilon_4 := -\nu\beta - \frac{\pi}{2}\left(\frac{1}{2}-\nu\right),$$

respectively, where $\beta := \arg((3+i\sqrt{4\lambda-9})/2\sqrt{\lambda})$. With (5.10) and its equivalent for t_4, and with (5.1), (5.8), and (5.9) we find the combined contribution from the third and fourth term of $g(t)$ to be

$$A2^{-k-\nu}(\sqrt{4\lambda-9})^{1/2-\nu}(\sqrt{\lambda})^{n+\nu}\left\{e^{i\varepsilon_3}\left(\frac{3+i\sqrt{4\lambda-9}}{2\sqrt{\lambda}}\right)^n + e^{i\varepsilon_4}\left(\frac{3-i\sqrt{4\lambda-9}}{2\sqrt{\lambda}}\right)^n\right\}$$

$$\sim \frac{n^{\nu-3/2}}{\Gamma(\nu)\sqrt{\pi}}(\sqrt{4\lambda-9})^{1/2-\nu}(\sqrt{\lambda})^{n+\nu}\cos\left\{\beta n+\nu\beta+\left(\frac{1}{4}-\frac{\nu}{2}\right)\pi\right\}.$$

This and (5.14) lead to Theorem 2(b).

(c) Let $\lambda = \frac{9}{4}$. Since $t_3 = t_4 = 1$, we have to find the coefficients of the expansion of

$$h(t) := F_k^{(1)}(t_2)(1-tt_1)^{1/2-\nu} + F_k^{(2)}(t_1)(1-tt_2)^{1/2-\nu} + F_k^{(5)}(1)(1-t)^{1-2\nu}.$$

The contribution from the first two terms of $h(t)$ is the same as in (5.14), with $\lambda = \frac{9}{4}$. Furthermore,

$$F_k^{(5)}(1) = \left(\frac{3}{4}\right)^{k-1/2} 2^{1/2-\nu}\Gamma \tag{5.15}$$

for $\lambda = \frac{9}{4}$, and the binomial theorem gives

$$(1-t)^{1-2\nu} = \sum_{n=0}^{\infty} \frac{\Gamma(2\nu-1+n)}{\Gamma(2\nu-1)n!}t^n. \tag{5.16}$$

Hence the contribution to the asymptotics of C_n from the third term of $h(t)$, with (5.16), (5.15), and (5.1) is

$$\frac{\Gamma(2\nu-1+n)}{n!\Gamma(2\nu-1)} \frac{\Gamma(k+\nu)\Gamma(\nu-1/2)}{\Gamma(\nu)\Gamma\left(\frac{k+\nu}{2}\right)\Gamma\left(\frac{k+\nu+1}{2}\right)}\left(\frac{3}{2}\right)^{n-1/2} 2^{1-k-\nu}. \tag{5.17}$$

By applying the duplication formula (5.12) twice, we get

$$\frac{\Gamma(k+\nu)\Gamma(\nu-1/2)}{\Gamma(2\nu-1)\Gamma\left(\frac{k+\nu}{2}\right)\Gamma\left(\frac{k+\nu+1}{2}\right)} = \frac{1}{\Gamma(\nu)}2^{1+k-\nu},$$

and Stirling's formula gives

$$\Gamma(2\nu-1+n)/n! \sim n^{2\nu-2} \qquad (\text{as } n\to\infty);$$

hence (5.17) is asymptotically equal to

$$\frac{1}{(\Gamma(\nu))^2}\left(\frac{n}{2}\right)^{2\nu-2}\left(\frac{3}{2}\right)^{n-1/2}.$$

This and (5.14) for $\lambda = \frac{9}{4}$ finally gives Theorem 2(c).

6. Further remarks. (1) Darboux's method can actually be used to find a complete asymptotic expansion for the $C_{n,k}^{\lambda,\nu}$; this would be a stronger result than Theorem 2. See, e.g., [5, Thm. 8.4].

(2) The sum of the elements of the nth row in the triangle (1.5) is easy to determine. Since the $C_{n,k}^{\lambda,\nu}$ are the coefficients of $f_n^{\lambda,\nu}(z)$, this sum is in fact $f_n^{\lambda,\nu}(1)$. Now (1.2) implies

$$f_n^{\lambda,\nu}(1) = (\sqrt{\lambda})^n C_n^{\nu}(3/2\sqrt{\lambda}).$$

More can be said in the case $\nu = 1$, since $C_n^1(x) = U_n(x)$. From (2.1) we get with a standard method (Binet's formula)

$$U_n(x) = \frac{1}{2\sqrt{x^2-1}}\{(x+\sqrt{x^2-1})^{n+1} - (x-\sqrt{x^2-1})^{n+1}\},$$

and therefore

$$f_n^{\lambda,1}(1) = \frac{(\sqrt{\lambda})^{n+1}}{\sqrt{9-4\lambda}}\left\{\left(\frac{3+\sqrt{9-4\lambda}}{2\sqrt{\lambda}}\right)^{n+1} - \left(\frac{3-\sqrt{9-4\lambda}}{2\sqrt{\lambda}}\right)^{n+1}\right\}. \tag{6.1}$$

As examples, we have

$$f_n^{0,1}(1) = 3^n,$$

$$f_n^{1,1}(1) = \frac{1}{\sqrt{5}}\left\{\left(\frac{3+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{3-\sqrt{5}}{2}\right)^{n+1}\right\}$$

(the odd-index Fibonacci numbers 1, 3, 8, 21, $\cdots$);

$$f_n^{2,1}(1) = 2^{n+1} - 1$$

(see (1.1)), and

$$f_n^{9/4,1}(1) = (n+1)\left(\frac{3}{2}\right)^n.$$

If $\lambda > \frac{9}{4}$ then (6.1) can be rewritten as

$$f_n^{\lambda,1}(1) = 2(4\lambda-9)^{-1/2}(\sqrt{\lambda})^{n+1}\sin\{(n+1)\theta\} \tag{6.2}$$

where θ is such that $\exp(i\theta) = (3+i\sqrt{4\lambda-9})/2\sqrt{\lambda}$, or $\theta = \mathrm{Cos}^{-1}(3/2\sqrt{\lambda})$. (6.2) gives easy explicit formulas for $\lambda = 3 (\theta = \pi/6)$, $\lambda = \frac{9}{2}$ $(\theta = \pi/4)$, $\lambda = 9$ $(\theta = \pi/3)$.

(3) The generating function $G^{\lambda,1}(z,t)$ of § 1 can be generalized as follows. Let $p(z) := a_0 + a_1 z + \cdots + a_r z^r$ and $q(z) = b_0 + b_1 z + \cdots + b_s z^s$ and expand

$$G(z,t) := \frac{1}{1 - tp(z) + t^2 q(z)} = \sum_{n=0}^{\infty} Q_n(z)(z)t^n.$$

Then $Q_0(z) = 1$, $Q_1(z) = p(z)$, and

$$Q_{n+1}(z) = p(z)Q_n(z) - q(z)Q_{n-1}(z), \tag{6.3}$$

and we see that $Q_n(z)$ is a polynomial of degree $\leqq nr$. If we denote

$$Q_n(z) = C_0^{(n)} + C_1^{(n)} z + \cdots + C_{nr}^{(n)} z^{nr},$$

we have the recursion

$$C_k^{(n+1)} = \sum_{j=0}^{k} a_j C_{k-j}^{(n)} - \sum_{j=0}^{k} b_j C_{k-j}^{(n-1)} \tag{6.4}$$

where $a_j := 0$ for $j < 0, j > r$, and $b_j := 0$ for $j < 0, j > s$. We note the following special cases.

(a) $p(z) := 1 + z$, $q(z) := 0$ gives $Q_n(z) = (1-z)^n$, and the $C_k^{(n)}$ are the binomial coefficients.

(b) $p(z) := 1 + z + z^2$, $q(z) = \lambda z^2$; this is the case dealt with in this paper, with $\nu = 1$.

(c) To generalize (b), we set $p(z) := 1 + z + \cdots + z^{2m}$, $q(z) := \lambda z^{2m}$. After reindexing (so that $Q_n(z) = C_{nm}^{(n)} + C_{nm-1}^{(n)} z + \cdots + C_0^{(n)} z^{nm} + \cdots + C_{nm-1}^{(n)} z^{2nm-1} + C_{nm}^{(n)} z^{2nm}$) (6.4) becomes

$$C_k^{(n+1)} = C_{k-m}^{(n)} + \cdots + C_k^{(n)} + \cdots + C_{k+m}^{(n)} - \lambda C_k^{(n-1)}; \tag{6.5}$$

this is the analogue to (1.6). No attempt has been made to determine the $C_k^{(n)}$ explicitly or asymptotically. However, it is easy to derive the sums of the elements in the rows of the triangle generated by (6.5). In analogy to and as a generalization of (6.1) we obtain

$$Q_n(1) = \frac{(\sqrt{\lambda})^{n+1}}{\sqrt{(2m+1)^2-4\lambda}} \cdot \left\{ \left(\frac{2m+1+\sqrt{(2m+1)^2-4\lambda}}{2\sqrt{\lambda}} \right)^{n+1} - \left(\frac{2m+1-\sqrt{(2m+1)^2-4\lambda}}{2\sqrt{\lambda}} \right)^{n+1} \right\}.$$

For $\lambda \leqq (2m+1)^2/4$, $Q_n(1)$ is positive for all n, and for $\lambda > (2m+1)^2/4$ it is an alternating sequence.

(4) K. B. Stolarsky [4] recently studied the recurrence $p_0(x) = 1$, $p_1(x) = x$, and

$$p_n(x) = x^n p_{n-1}(x^{-1}) + p_{n-2}(x).$$

He showed that for $n \geqq 0$

$$p_{2n+1}(x) = x f_{2n+1}^{1,1}(x)$$

(in our notation), i.e., the $p_{2n+1}(x)$ are self-inverse polynomials, or in other words, the coefficients are "centrally symmetric." However, it is easy to see that the $p_{2n}(x)$ do not have this property; in fact, it is shown in [4] that the coefficients of $p_{2n}(x)$ are "strongly noncentrally symmetric."

Acknowledgments. I wish to thank Kenneth B. Stolarsky for suggesting the problem that led to this paper, and for useful discussions. I also thank the referee for suggesting various improvements.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, 1970.

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

[3] E. D. RAINVILLE, *Special Functions*, MacMillan, New York, 1960.

[4] K. B. STOLARSKY, *A recurrence with deviating arguments that generates intermittent symmetry*, preprint.

[5] G. SZEGÖ, *Orthogonal Polynomials*, 4th ed., American Mathematical Society, Providence, RI, 1975.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

016

A QUANTITATIVE DIRICHLET–JORDAN TYPE THEOREM FOR ORTHOGONAL POLYNOMIAL EXPANSIONS*

H. N. MHASKAR†

Abstract. In 1985, Bojanic estimated the rate at which the partial sums of the orthogonal expansion of a function of bounded variation converge to the function. We generalize and sharpen his estimates when the function being expanded is an iterated integral of a function of bounded variation. The class of orthogonal polynomials considered include the Jacobi polynomials orthogonal on $[-1, 1]$ with respect to $(1-x)^\alpha(1+x)^\beta$, $\alpha, \beta \geqq -\frac{1}{2}$. For the Jacobi series, our method gives an asymptotic expression for the difference between the function and its partial sums at points of discontinuity of a high order derivative of the function. In particular, discontinuities in an even order derivative are shown not to affect the rate of convergence.

Key words. orthogonal series, Jacobi polynomials, bounded variation

AMS(MOS) subject classifications. 41A25, 42C15, 33A65

The well-known Dirichlet-Jordan convergence criterion for Fourier series states that the trigonometric Fourier series of a 2π-periodic function f having bounded variation converges to $\frac{1}{2}[f(x+0)+f(x-0)]$ for every x and this convergence is uniform on every closed interval on which f is continuous [13, Thm. 2.8.1].

Many mathematicians have studied the generalizations and analogues of this criterion for series other than Fourier series, especially orthogonal polynomial series ([9], [12], [5], [11], [6], [10]). The most general theorem known to us about orthogonal polynomial series on $[-1, 1]$ is due to Freud [5].

To describe Freud's theorem, we develop some notation.

A function $w: [-1, 1] \to [0, \infty)$ is called a weight function if

$$\int_{-1}^{1} |t|^k w(t)\, dt < \infty, \qquad k = 0, 1, 2, \cdots. \tag{1}$$

The class of all polynomials of degree not exceeding n will be denoted by Π_n. When w is a weight function, there is a unique system $\{p_n\}$ of polynomials such that

$$p_n(w, x) := p_n(x) := \gamma_n x^n + \cdots \in \Pi_n, \qquad \gamma_n > 0 \tag{2}$$

and

$$\int_{-1}^{1} p_n(x) p_m(x) w(x)\, dx = \delta_{nm}. \tag{3}$$

If f is integrable on $[-1, 1]$, we put

$$c_k := c_k(w, f) := \int_{-1}^{1} f(t) p_k(t) w(t)\, dt, \qquad k = 0, 1, 2, \cdots, \tag{4a}$$

$$s_m(w, f, x) := s_m(f, x) := \sum_{k=0}^{m-1} c_k p_k(x), \qquad m = 1, 2, \cdots. \tag{4b}$$

Freud's theorem now states the following.

THEOREM 1 [5, Thm. V.7.5]. *Suppose w is a weight function satisfying*

$$0 < w(x) \leqq M(1-x^2)^{-1/2}, \qquad x \in (-1, 1). \tag{5}$$

* Received by the editors February 18, 1986; accepted for publication February 9, 1987.

† Department of Mathematics and Computer Science, California State University, Los Angeles, California 90032.

Further, let f be a function of bounded variation on $[-1, 1]$. Then $s_n(w, f, x)$ converges to $\frac{1}{2}(f(x+0)+f(x-0))$ for every $x \in (-1, 1)$ where w is continuous; the convergence is uniform in every $[a, b] \subset (-1, 1)$ on which f and w are continuous.

In particular, Freud's theorem applies to all the Jacobi polynomial expansions for $w(x) := (1-x)^{\alpha}(1+x)^{\beta}$ and $\alpha, \beta \geqq -\frac{1}{2}$. However, despite the applications of such theorems in numerical analysis and differential equations [8], only a few results are known about the rate at which the convergence takes place. In the case of classical Fourier series, one such theorem, unimprovable in some sense, was proved by Bojanic [1]. For orthogonal polynomial expansions, he proved the following.

THEOREM 2 ([2]). *Let w be a weight function and suppose that for $x \in (-1, 1)$ and $n = 1, 2, \cdots$*

$$0 < w(x) \leqq M(1-x^2)^{-A}, \tag{6}$$

$$|p_n(x)| \leqq M(1-x^2)^{-B}, \tag{7}$$

$$\left| \int_{-1}^{x} w(t)p_n(t)\, dt \right| \leqq \frac{C}{n}. \tag{8}$$

If f is a function of bounded variation on $[-1, 1]$ then

$$\begin{aligned} |s_n(w, f, x) - \tfrac{1}{2}(f(x+)+f(x-))| \leqq \frac{C(x)}{n} \sum_{k=1}^{n} V\left(\left[x - \frac{1+x}{k}, x + \frac{1-x}{k}\right], \bar{g}_x\right) \\ + \tfrac{1}{2}|f(x+)-f(x-)| \cdot |s_n(w, \psi_x, x)| \end{aligned} \tag{9}$$

where $\bar{g}_x$ is given by

$$\bar{g}_x(t) := \begin{cases} f(t)-f(x-), & -1 \leqq t < x, \\ 0, & t = x, \\ f(t)-f(x+), & x < t \leqq 1, \end{cases} \tag{10}$$

and

$$\psi_x(t) := \operatorname{sign}(t-x) := \begin{cases} \dfrac{t-x}{|t-x|}, & t \neq x, \\ 0, & t = x. \end{cases} \tag{11}$$

Moreover, $C(x) > 0$ for $x \in (-1, 1)$ and $V([a, b], g)$ is the total variation of g on $[a, b]$.

Under the conditions of Theorem 1, $\{S_n(\psi_x, x)\}$ tends to zero as $n \to \infty$. In the case of Jacobi expansions, it was noted in [2] that

$$|S_n(\psi_x, x)| \leqq \frac{M(x)}{n}. \tag{12}$$

We do not expect (12) to hold for a general weight function. It is possible to give an analogue where $1/n$ is replaced by a relatively complicated expression involving quantities related to w and x alone and we hope to obtain "easier" estimates on this expression in future.

While the techniques used in [2] are powerful enough to yield similar theorems in a more general setting than orthogonal polynomials, it is extremely difficult to extend them to study the case when the function has higher derivatives. Thus, for instance, the results in [3] which apply only to the special case of Legendre polynomial expansions are far more complicated than Theorem 2 and, in retrospect, require undue restrictions on the function to get a "good" rate.

In this paper, we study the case when the function has higher order derivatives and obtain a rate of convergence theorem under general conditions similar to those in Theorem 2. For the case of the Jacobi polynomial expansions, we get, in fact, an asymptotic expression for $s_n(w, f, x)$ when the high order derivative of f has a discontinuity at x. More precisely, we prove the following theorem in the general case.

THEOREM 3. *Suppose w is a weight function and*

$$|p_n(w, x)w(x)\sqrt{1-x^2}| \leqq M, \qquad x \in [-1, 1]. \tag{13}$$

Further suppose that $r \geqq 1$ is an integer, f is a $(r-1)$-times continuously differentiable function on $[-1, 1]$ and, for $x \in [-1, 1]$

$$f^{(r-1)}(x) = f^{(r-1)}(-1) + \int_{-1}^{x} \phi(t)\, dt \tag{14}$$

where ϕ is a function of bounded variation on $[-1, 1]$.

Then, for $x \in (-1, 1)$ and $n \geqq (1-x^2)^{-1}$,

$$\begin{aligned} &\left| s_n(w, f, x) - f(x) - \frac{1}{r!}[\phi(x+) - \phi(x-)] \cdot s_n(w, \Gamma_r(x, \cdot), x) \right| \\ &\leqq \frac{c_1}{(1-x^2)^{3/2} w(x)} \cdot \frac{1}{n^{r+1}} \sum_{k=1}^{n} V\left(\left[x - \frac{1+x}{n}, x + \frac{1-x}{n}\right], g_x\right) \end{aligned} \tag{15}$$

where

$$\Gamma_r(x, t) := \begin{cases} 0 & \text{if } t \leqq x, \quad -1 \leqq t \leqq x, \\ (t-x)^r & \text{if } t > x, \quad x < t \leqq 1, \end{cases} \tag{16}$$

and

$$g_x(t) := \begin{cases} \phi(t) - \phi(x-), & t < x, \\ 0, & t = x, \\ \phi(t) - \phi(x+), & t > x. \end{cases} \tag{17}$$

Here and in the sequel, $c, c_1, c_2, \cdots$, will denote constants whose value may be different in different occurrences of the symbol, even within a single formula. Constants which will retain their values will be denoted by capital letters.

In the case of the Jacobi weights $w(x) := w_{\alpha,\beta}(x) := (1-x)^{\alpha}(1+x)^{\beta}$, $\alpha, \beta \geqq -\frac{1}{2}$ we can explicitly compute $s_n(w_{\alpha,\beta}, \Gamma_r(x, \cdot), x)$ in terms of the Jacobi polynomials and hence, obtain asymptotics for the same. This is summarized in the following.

THEOREM 4. *We have, for $x \in (-1, 1)$,*

$$\left| s_n(w_{\alpha,\beta}, \Gamma_r(x, \cdot), x) - \frac{(1-x^2)^{r/2} \sin(r\pi/2)}{n^r \cdot \pi}(r-1)! \right| \leqq \frac{c_2(\alpha, \beta, x)}{n^{r+1}}.$$

The proof of Theorem 3, similar to the ones in [1]-[3], is based on a repeated integration by parts. However, we make a more effective use of the orthogonality of p_n's. In particular, instead of having to estimate the integrals involving the "Christoffel-Darboux kernel" as in [2], [3], we can apply the results of G. Freud on the one-sided L^1-approximation [7] of the functions Γ_r defined in (16).

To prove Theorem 3, we start with the formula

$$f(t) = f(x) + P(x, t) + \frac{1}{(r-1)!} \int_x (t-u)^{r-1} \phi(u)\, du$$

where $P(x, \cdot) \in \Pi_r$ and $P(x, x) = 0$. This formula can be easily verified using Taylor's theorem with the integral form for the remainder and (14). A simple computation now gives, for $t \neq x$,

$$f(t) = f(x) + P(x, t) + \frac{[\phi(x+) - \phi(x-)]}{r!}\Gamma_r(x, t) + F(x, t) \tag{18}$$

where

$$F(x, t) := \frac{1}{(r-1)!}\int_x^t (t-u)^{r-1} g_x(u)\, du = \frac{1}{r!}\int_x^t (t-u)^r\, dg_x(u). \tag{19}$$

Now,

$$s_n(w, P(x, \cdot), x) = P(x, x) = 0 \tag{20}$$

and hence,

$$\left| s_n(w, f, x) - f(x) - \frac{[\phi(x+) - \phi(x-)]}{r!} s_n(w, \Gamma_r(x, \cdot), x) \right| = |s_n(w, F(x, \cdot), x)|. \tag{21}$$

For simplicity of notation, we shall write in the sequel, g instead of g_x and $F(t)$ instead of $f(x, t)$. The following lemma summarizes certain technical estimates which will be needed, especially, to estimate the integrated terms when we use the integration by parts formula repeatedly.

LEMMA 5. *Put*

$$G(t) := \frac{(x-t)^{-1}}{r!}\int_x^t (t-u)^r\, dg(u) = (x-t)^{-1} F(t), \tag{22}$$

$$\Lambda_n(t) := \frac{1}{r!}\int_{-1}^t (t-u)^r p_n(u) w(u)\, du. \tag{23}$$

Then for integer k, $0 \leqq k \leqq r$,

$$|G^{(k)}(t)| \leqq c \cdot |x-t|^{r-k-1} \left| \int_x^t |dg(u)| \right|, \tag{24}$$

$$|\Lambda_n^{(k)}(t)| \leqq c \cdot n^{k-r-1}. \tag{25}$$

Proof. Estimate (24) follows easily when we compute $G^{(k)}(t)$ using Leibnitz's rule. Observe that

$$\Lambda_n^{(k)}(t) = \frac{1}{(r-k)!}\int_{-1}^t (t-u)^{r-k} p_n(u) w(u)\, du.$$

Thus,

$$\begin{aligned} \Lambda_n^{(k)}(t) &= \frac{1}{(r-k)!}\int_{-1}^1 \Gamma_{r-k}(u, t) p_n(u) w(u)\, du \\ &= \frac{1}{(r-k)!}\int_{-1}^1 [\Gamma_{r-k}(u, t) - P(u)] p_n(u) w(u)\, du \end{aligned} \tag{26}$$

for an arbitrary polynomial $P \in \Pi_{n-1}$. Hence, using (13),

$$|\Lambda_n^{(k)}(t)| \leqq c \cdot \inf_{P \in \Pi_{n-1}} \int_{-1}^1 |\Gamma_{r-k}(u, t) - P(u)| \frac{du}{\sqrt{1-u^2}}. \tag{27}$$

Estimate (25) now follows from formulas (15) and (18) of [7, pp. 15-16]. □

Proof of Theorem 3. In a view of (21), it is enough to estimate $|s_n(w, F, x)|$, when $n \geqq (1-x^2)^{-1}$.

It is well known [11, Thm. 3.2.2] that

$$s_n(w, F, x) = \int_{-1}^{1} K_n(x, t)F(t)w(t)\,dt \tag{28}$$

where, with the notation of (2),

$$K_n(x, t) := \frac{\gamma_{n-1}}{\gamma_n}\,\frac{P_n(x)p_{n-1}(t) - p_n(t)p_{n-1}(x)}{x - t}. \tag{29}$$

Let $\underline{x} := x - (1+x)/n$, $\bar{x} := x + (1-x)/n$. Then,

$$s_n(w, F, x) = A + B + C \tag{30}$$

where

$$A := \int_{-1}^{\underline{x}} K_n(x, t)F(t)w(t)\,dt, \tag{31a}$$

$$B := \int_{\underline{x}}^{\bar{x}} K_n(x, t)F(t)w(t)\,dt, \tag{31b}$$

$$C := \int_{\bar{x}}^{1} K_n(x, t)F(t)w(t)\,dt. \tag{31c}$$

In view of (13),

$$|K_n(x, t)| = \left|\sum_{k=0}^{n-1} p_k(x)p_k(t)\right| \leqq \frac{c}{w(x)\sqrt{1-x^2}} \cdot \frac{n}{w(t)\sqrt{1-t^2}}. \tag{32}$$

So

$$|B| \leqq c.n.[w(x)\sqrt{1-x^2}]^{-1} \cdot \int_{\underline{x}}^{\bar{x}} \left(\int_{x}^{t} (t-u)^r |dg(u)|\right) \cdot \frac{dt}{\sqrt{1-t^2}}. \tag{33}$$

$$\leqq c[w(x)\sqrt{1-x^2}]^{-1} n^{-r+1} V([\underline{x}, \bar{x}], g) \cdot \int_{\underline{x}}^{\bar{x}} \frac{dt}{\sqrt{1-t^2}}.$$

Since $n \geqq (1-x^2)^{-1}$, it is easy to see that

$$\int_{\underline{x}}^{\bar{x}} \frac{dt}{\sqrt{1-t^2}} \leqq \frac{c}{n\sqrt{1-x^2}}.$$

So,

$$|B| \leqq c \cdot [w(x)(1-x^2)]^{-1} \cdot n^{-r} V([\underline{x}, \bar{x}], g). \tag{34}$$

In view of (29), (22),

$$A = \int_{-1}^{\underline{x}} \frac{\gamma_{n-1}}{\gamma_n}[p_n(x)p_{n-1}(t) - p_{n-1}(x)p_n(t)]G(t)w(t)\,dt$$

$$= \frac{\gamma_{n-1}}{\gamma_n}\left\{p_n(x) \cdot \int_{-1}^{\underline{x}} \Lambda_{n-1}^{(r+1)}(t)G(t)\,dt - p_{n-1}(x) \int_{-1}^{\underline{x}} \Lambda_n^{(r+1)}(t)G(t)\,dt\right\}. \tag{35}$$

Thus, using (13), and the following estimate

$$\frac{\gamma_{n-1}}{\gamma_n} = \int_{-1}^{1} xp_{n-1}(u)p_n(u)w(u)\,du \leqq 1$$

(which can be seen easily using the Schwarz inequality), we get

$$|A| \leqq c[w(x)\sqrt{1-x^2}]^{-1}\{|A_1| + |A_2|\} \tag{36}$$

where

$$A_1 := \int_{-1}^{\underline{x}} \Lambda_{n-1}^{(r+1)}(t) G(t)\, dt, \tag{37a}$$

$$A_2 := \int_{-1}^{\underline{x}} \Lambda_{n}^{(r+1)}(t) G(t)\, dt. \tag{37b}$$

We estimate A_2, the estimate for A_1 being similar. Integrating by parts several times, we obtain

$$A_2 = \sum_{j=0}^{r} (-1)^j \Lambda_n^{(r-j)}(\underline{x}) G^{(j)}(\underline{x}) + (-1)^{r+1} \int_{-1}^{\underline{x}} \Lambda_n(t)\, dG^{(r)}(t). \tag{38}$$

Using Lemma 5 and then Leibnitz's rule, we obtain

$$\begin{aligned} |A_2| &\leqq \frac{c}{n^r} V([\underline{x}, x], g) + \frac{c}{n^{r+1}} \int_{-1}^{\underline{x}} |dG^{(r)}(t)| \\ &\leqq \frac{c}{n^r} V([\underline{x}, x], g) + \frac{c}{n^{r+1}} \int_{-1}^{\underline{x}} \frac{V(t)}{(x-t)^2}\, dt + \frac{c}{n^{r+1}} \int_{-1}^{\underline{x}} \frac{dV(t)}{x-t} \end{aligned} \tag{39}$$

where

$$V(t) := V([t, x], g).$$

Integrating by parts once more in the last integral, we obtain

$$\begin{aligned} |A_2| &\leqq \frac{c}{n^{r+1}} \cdot nV(\underline{x}) + \frac{c}{n^{r+1}} \int_{-1}^{\underline{x}} \frac{V(t)}{(x-t)^2}\, dt \\ &\leqq \frac{c}{n^{r+1}} \int_{-1}^{\underline{x}} \frac{V(t)}{(x-t)^2}\, dt. \end{aligned} \tag{40}$$

Similarly,

$$|A_1| \leqq \frac{c}{(n-1)^{r+1}} \int_{-1}^{\underline{x}} \frac{V(t)}{(x-t)^2}\, dt \leqq \frac{c}{n^{r+1}} \int_{-1}^{\underline{x}} \frac{V(t)}{(x-t)^2}\, dt. \tag{41}$$

Since $V(t)$ is decreasing, it follows from (40), (41), (36) that

$$\begin{aligned} |A| &\leqq c[w(x)\sqrt{1-x^2}]^{-1}(1+x)^{-1} \frac{1}{n^{r+1}} \sum_{k=1}^{n} V\left(\left[x - \frac{1+x}{k}, x\right], g\right) \\ &\leqq c[w(x)(1-x^2)^{3/2}]^{-1} \frac{1}{n^{r+1}} \sum_{k=1}^{n} V\left(\left[x - \frac{1+x}{k}, x\right], g\right). \end{aligned} \tag{42}$$

Similarly,

$$|C| \leqq c[w(x)(1-x^2)^{3/2}]^{-1} \frac{1}{n^{r+1}} \sum_{k=1}^{n} V\left(\left[x, x + \frac{1+x}{k}\right], g\right). \tag{43}$$

Substituting from (43), (42), (34) into (30), we see that

$$|s_n(w, F, x)| \leqq c \cdot [w(x)(1-x^2)^{3/2}]^{-1} \frac{1}{n^{r+1}} \sum_{k=1}^{n} V\left(\left[x - \frac{1+x}{k}, x + \frac{1-x}{k}\right], g\right). \tag{44}$$

In view of (21), this completes the proof of Theorem 3. □

Proof of Thoerem 4. The proof relies heavily on the special properties of the Jacobi polynomials. For the convenience of the reader, therefore, we use the standard notation as in [11] or [4]. Thus, with $w(x) := (1-x)^\alpha(1+x)^\beta$

$$P_n^{(\alpha,\beta)}(x) := \{h_n^{(\alpha,\beta)}\}^{1/2} p_n(w, x) \tag{45}$$

where

$$h_b^{(\alpha,\beta)} := \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \cdot \frac{\Gamma(n+\alpha+1)(n+\beta+1)}{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}. \tag{46}$$

Using Stirling's approximation, we see that

$$h_n^{(\alpha,\beta)} = \frac{2^{\alpha+\beta}}{n}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{47}$$

Moreover, the Christoffel-Darboux kernel in (28), (29) is given by the equation 4.5.2 in [11]

$$K_n(x, t) = A_n \cdot \frac{P_n^{(\alpha,\beta)}(t)P_{n-1}^{(\alpha,\beta)}(x) - P_{n-1}^{(\alpha,\beta)}(t)P_n^{(\alpha,\beta)}(x)}{t-x} \tag{48a}$$

where

$$A_n := \frac{2^{-\alpha-\beta}}{2n+\alpha+\beta} \frac{\Gamma(n+1)\Gamma(n+\alpha+\beta+1)}{\Gamma(n+\alpha)\Gamma(n+\beta)} = 2^{-\alpha-\beta-1} n\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{48b}$$

Thus,

$$s_n(w, \Gamma_r(x, \cdot), x) = A_n\{P_{n-1}^{(\alpha,\beta)}(x)I_n(x) - P_n^{(\alpha,\beta)}(x)I_{n-1}(x)\} \tag{49}$$

where

$$I_n(x) := \int_x^1 (t-x)^{r-1} P_n^{(\alpha,\beta)}(t)(1-t)^\alpha(1+t)^\beta\, dt. \tag{50}$$

Now, the Rodrigues' formula [11, eq. (4.3.1)] implies (cf. [4, eq. 10.8(38)]) that

$$\int_x^1 P_n^{(\alpha,\beta)}(t)(1-t)^\alpha(1+t)^\beta\, dt = \frac{(1-x)^{\alpha+1}(1+x)^{\beta+1}P_{n-1}^{(\alpha+1,\beta+1)}(x)}{2n}. \tag{51}$$

Applying (51) repeatedly r times, we see that

$$I_n(x) = \frac{(r-1)!(n-r)!}{2^r n!}(1-x)^{\alpha+r}(1+x)^{\beta+r}P_{n-r}^{(\alpha+r,\beta+r)}(x). \tag{52}$$

Evaluating $I_{n-1}(x)$ in the same way and then substituting into (49), we get, after a little bit of simplification,

$$\begin{aligned} s_n(w, \Gamma_r(x, \cdot), x) = {} & \frac{(n-1-r)!}{2^r(n-1)!}A_n(1-x)^{\alpha+r}(1+x)^{\alpha+r}(1+x)^{\beta+r}(r-1)! \\ & \cdot \{P_{n-1}^{(\alpha,\beta)}(x)P_{n-r}^{(\alpha+r,\beta+r)}(x) - P_n^{(\alpha,\beta)}(x)P_{n-1-r}^{(\alpha+r,\beta+r)}(x)\} \\ & - \frac{(n-r-1)!r!}{2^r n!}A_n(1-x)^{\alpha+r}(1+x)^{\beta+r}P_{n-1}^{(\alpha,\beta)}(x)P_{n-r}^{(\alpha+r,\beta+r)}(x). \end{aligned} \tag{53}$$

Now, according to the Darboux formula, [11, Thm. 8.21.8], we have, with $x := \cos\theta$,

$$\begin{aligned} (1-x)^{\alpha/2}(1+x)^{\beta/2}P_n^{(\alpha,\beta)}(x) = {} & n^{-1/2}\pi^{-1/2}(1-x^2)^{-1/4}2^{(\alpha+\beta+1)/2} \\ & \cdot \cos\left[\left(n+\frac{\alpha+\beta+1}{2}\right)\theta - (\alpha+\tfrac{1}{2})\frac{\pi}{2}\right] + \mathcal{O}(n^{-3/2}) \quad \text{if } x \in (-1, 1). \end{aligned} \tag{54}$$

(Here, and in the sequel, the big $\mathcal{O}$ term may depend upon x.) In particular,

$$(1-x)^{\alpha/2}(1+x)^{\beta/2}P_n^{(\alpha,\beta)}(x)=\mathcal{O}(n^{-1/2}). \tag{55}$$

If we use (55) and (48b), a few simple computations reduce (53) to

$$\begin{aligned} s_n(w,\Gamma_r(x,\cdot),x)=&\frac{(r-1)!2^{-\alpha-\beta-1}}{2^r n^{r-1}}\\ &\cdot\{P_{n-1}^{(\alpha,\beta)}(x)P_{n-r}^{(\alpha+r,\beta+r)}(x)-P_n^{(\alpha,\beta)}(x)P_{n-r-1}^{(\alpha+r,\beta+r)}(x)\}\\ &\cdot(1-x)^{\alpha+r}(1+x)^{\beta+r}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)+\mathcal{O}\left(\frac{1}{n^{r+1}}\right). \end{aligned} \tag{56}$$

Further, using (54), we get

$$\begin{aligned} &(1-x)^{\alpha+r}(1+x)^{\beta+r}P_{n-1}^{(\alpha,\beta)}(x)P_{n-r}^{(\alpha+r,\beta+r)}(x)\\ &=(1-x)^{\alpha/2}(1+x)^{\beta/2}P_{n-1}^{(\alpha,\beta)}(x)(1-x)^{(\alpha+r)/2}(1+x)^{(\beta+r)/2}P_{n-r}^{(\alpha+r,\beta+r)}(x)(1-x^2)^{r/2}\\ &=(1-x^2)^{r/2}n^{-1}\pi^{-1}(1-x^2)^{-1/2}2^{\alpha+\beta+1}2^r\\ &\quad\cdot\cos\left[\left(n-1+\frac{\alpha+\beta+1}{2}\right)\theta-\left(\alpha+\frac{1}{2}\right)\frac{\pi}{2}\right]\\ &\quad\cdot\cos\left[\left(n-r+\frac{\alpha+\beta+1+2r}{2}\right)\theta-\left(\alpha+r+\frac{1}{2}\right)\frac{\pi}{2}\right]+\mathcal{O}(n^{-2})\\ &=(1-x^2)^{r/2}n^{-1}\pi^{-1}(1-x^2)^{-1/2}2^{\alpha+\beta}2^r\\ &\quad\cdot\left\{\cos\left[(2n+\alpha+\beta)\theta-(2\alpha+r+1)\frac{\pi}{2}\right]+\cos\left(\theta-\frac{r\pi}{2}\right)\right\}+\mathcal{O}(n^{-2}). \end{aligned} \tag{57}$$

Similarly,

$$\begin{aligned} &(1-x)^{\alpha+r}(1+x)^{\beta+r}P_n^{(\alpha,\beta)}(x)P_{n-r-1}^{(\alpha+r,\beta+r)}(x)\\ &=(1-x^2)^{r/2}n^{-1}\pi^{-1}(1-x^2)^{-1/2}2^{\alpha+\beta}2^r\\ &\quad\cdot\left\{\cos\left[(2n+\alpha+\beta)\theta-(2\alpha+r+1)\frac{\pi}{2}\right]+\cos\left(\theta+\frac{r\pi}{2}\right)\right\}+\mathcal{O}(n^{-2}). \end{aligned} \tag{58}$$

Substituting from (57), (58) into (56), we get

$$\begin{aligned} s_n(w,\Gamma_r(x,\cdot),x)=&\frac{(r-1)!}{2^r n^{r-1}}2^{-\alpha-\beta-1}(1-x^2)^{r/2}n^{-1}\pi^{-1}(1-x^2)^{-1/2}2^{\alpha+\beta}2^r\\ &\cdot\left\{\cos\left(\theta-\frac{r\pi}{2}\right)-\cos\left(\theta+\frac{r\pi}{2}\right)\right\}+\mathcal{O}\left(\frac{1}{n^{r+1}}\right)\\ =&\frac{(r-1)!(1-x^2)^{r/2}}{\pi n^r}\sin\frac{r\pi}{2}\sin\theta(1-x^2)^{-1/2}+\mathcal{O}\left(\frac{1}{n^{r+1}}\right)\\ =&\frac{(r-1)!}{\pi n^r}\sin\frac{r\pi}{2}(1-x^2)^{r/2}+\mathcal{O}\left(\frac{1}{n^{r+1}}\right). \end{aligned} \tag{59}$$

□

In the general case, we do not yet have the analogue of Theorem 4, but using condition (13) and the one-sided approximation rsesults of Freud in [7], it is not hard to see that

$$|s_n(w, \Gamma_r(x, \cdot), x)| \leqq \frac{C(x)}{n^r}. \tag{60}$$

The ideas in the proof of Theorem 3 can be extended to obtain a similar rate of convergence for expansions in polynomials orthogonal on the whole real axis with respect to what is now known as the Freud weights. This, however, will be discussed in separate papers.

Acknowledgments. The author wishes to thank Professor Ranko Bojanic (Ohio State University, Columbus, Ohio) for his generous help in the research presented in this paper. The author is grateful to The Bowling Green State University, Bowling Green, Ohio for their support while this work was being done.

REFERENCES

[1] R. BOJANIC, *An estimate of the rate of convergence for Fourier series of functions of bounded variation*, Publ. Inst. Math. (Belgrade), 26 (40) (1979), pp. 57-60.

[2] ———, *An estimate for the rate of convergence of a general class of orthogonal polynomial expansions of functions of bounded variation*, manuscript.

[3] R. BOJANIC AND Z. DIVIS, *Asymptotic behavior of partial sums of Fourier-Legendre series*, manuscript.

[4] A. ERDELYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[5] G. FREUD, *Orthogonal Polynomials*, Pergamon Press, Elmsford, NY, 1971.

[6] ———, *Extension of the Dirichlet-Jordan criterion to a general class of orthogonal polynomial expansions*, Acta Math. Hungar., 25 (1974), pp. 109-122.

[7] ———, *Über Einseitige Approximation durch Polynome*, I, Acta Sci. Math. (Szeged), 16 (1955), pp. 12-28.

[8] D. GOTTLIEB AND S. A. ORSAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conference Series 26, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[9] G. H. HARDY AND J. E. LITTLEWOOD, *A convergence theorem for Fourier series*, Math. Z., 28 (1928), pp. 612-634.

[10] H. N. MHASKAR, *Extensions of the Dirichlet-Jordan criterion*, J. Approx. Theory, 42 (1984), pp. 138-148.

[11] G. SZEGÖ, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., 23, Providence, RI, 1978.

[12] E. C. TITCHMARSCH, *Eigenfunction Expansions Associated with Second Order Differential Equations*, I, Clarendon Press, Oxford, 1962.

[13] A. ZYGMUND, *Trigonometric Series*, Cambridge University Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 19, No. 2, March 1988

017

DETERMINANTS OF LAPLACIANS AND MULTIPLE GAMMA FUNCTIONS*

ILAN VARDI†

Abstract. In this paper we generalize the classical formula $\Gamma(\frac{1}{2})=\sqrt{\pi}$. We do this by recalling the Multiple Gamma Function first studied in the nineteenth century by Barnes and others. These functions at $\frac{1}{2}$ will be expressed in terms of the functional determinant of Laplacians of the n-sphere (thus these invariants of the n-sphere generalize π). Determinants of Laplacians have been a recent subject of research due to their relevance to Superstring Theory. While the determinant of the Laplacian has been computed for a flat torus using the Kronecker Limit Formula, our result gives the case of the n-sphere with the standard metric.

1. Introduction.

In this paper we generalize the classical formula

$$\Gamma(\tfrac{1}{2})=\sqrt{\pi}. \tag{1}$$

To do so we reinterpret (1) using the functional determinant of the Laplacian on a compact manifold.

Let Δ be the Laplacian of the compact manifold M. Then Δ has a discrete sequence of eigenvalues

$$0=\lambda_0<\lambda_1\leqq\lambda_2\leqq\cdots,$$

which, by the so-called Weyl Law have an asymptotic formula for λ_n as $n\to\infty$. Using this fact one can show that

$$F(s)=\sum_1^\infty \frac{1}{\lambda_n^s}$$

converges absolutely in a half plane $\operatorname{Re}(s)>\sigma$. Thus

$$F'(s)=-\sum_{n=1}^{\infty}\frac{\log\lambda_n}{\lambda_n^s},\qquad \operatorname{Re}(s)>\sigma.$$

One sees that *formally* $e^{-F'(0)}$ is the product of the eigenvalues of Δ. This product does not converge, but it can be shown that $F(s)$ can be continued analytically to $s=0$ and we define

$$\det\Delta=e^{-F'(0)}$$

to be the *Functional Determinant* of Δ. It is easily shown that for $M=S^1$, the unit circle, with standard Laplacian $\Delta_1=d^2/dx^2$

$$\det\Delta_1=4\pi^2. \tag{2}$$

Thus (1) can be rewritten as

$$\Gamma(\tfrac{1}{2})=(\det\Delta_1)^{1/4}2^{-1/2}. \tag{3}$$

* Received by the editors September 22, 1986; accepted for publication (in revised form) February 9, 1987.

† Department of Mathematics, Stanford University, Stanford, California 94305.

We generalize this formula to *Multiple Gamma Functions.* These are given by

(i) $\Gamma_0(x) = x^{-1}$,

(ii) $\Gamma_n(1) = 1$,

(iii) $\Gamma_{n+1}(x+1) = \Gamma_{n+1}(x)/\Gamma_n(x)$,

(iv) $\Gamma_n(x)^{-1}$ *is* C^∞ *on* $\mathbf{R}$,

(v) $(-1)^n(d^{n+1}/dx^{n+1}) \log \Gamma_n(x) \geqq 0$ for $x > 0$.

As in Vignéras [12] this defines $\Gamma_n(x)$ uniquely. For example

$$\Gamma_1(x) = \Gamma(x)$$

is the *Bohr-Mollerup Theorem* [1].

The Multiple Gamma Functions were first defined in a slightly different form by Barnes [2]. Our definition is due to Vignéras and seems more natural than that of Barnes.

$\Gamma_2(x)$ is the *Double Gamma Function,* which was originally studied in the latter half of the nineteenth century, notably by Barnes [3]. It has since been studied by Shintani in a similar form [10], while the function considered here was shown by Vignéras to occur naturally as the factor at infinity of the Selberg Zeta Function [12]. This has since been exploited by Sarnak [9].

Properties of the Double Gamma Function are:

(i) Weierstrass product:

$$\Gamma_2(x+1) = (2\pi)^{-x/2} \exp\left(\frac{x}{2} + \frac{\gamma+1}{2}x^2\right)\left(\prod_{k=1}^{\infty}\left(1+\frac{x}{k}\right)\exp\left(-x+\frac{x^2}{2k}\right)\right)^{-1}.$$

(ii) Stirling formula:

$$\log \Gamma_2(x+a+1) = -\frac{(x+a)}{2}\log 2\pi + \log A - \frac{1}{12} + \frac{3x^2}{4} + ax - \left(\frac{x^2}{2} - \frac{1}{12} + \frac{a^2}{2} + ax\right)\log x + \mathcal{O}\left(\frac{1}{x}\right)$$

where A is a constant of Kinkelin [5]

$$\log A = \lim_{k\to\infty}\left[\log(1^1 2^2 \cdots k^k) - \left(\frac{k^2}{2} + \frac{k}{2} + \frac{1}{2}\right)\log k + \frac{k^2}{4}\right].$$

In § 4 we will evaluate this as

$$\log A = \exp\left(\tfrac{1}{12} - \zeta'(-1)\right)$$

where $\zeta(s)$ is the Riemann ζ-function.

(iii) $\Gamma_2(n+2) = (1!2!\cdots n!)^{-1}$.

(iv) Multiplication formula:

$$\prod_{r=0}^{k-1}\prod_{s=0}^{k-1}\Gamma_2\left(x+\frac{r+s}{k}\right) = K(2\pi)^{-k(k-1)x/2}k^{(k^2x^2/2)-kx}\Gamma_2(kx)$$

where

$$K = A^{k^2-1}\, e^{(1-k^2)/12}(2\pi)^{(k-1)/2}k^{5/12}.$$

This formula, for the case $k = 2$, will be proved in § 2.

(v) Generalized Hilbert Determinant:

$$\begin{vmatrix} \frac{1}{1+1+b} & \frac{1}{1+2+b} & \cdots & \frac{1}{1+n+b} \\ \frac{1}{2+1+b} & \frac{1}{2+2+b} & \cdots & \frac{1}{2+n+b} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{n+1+b} & \frac{1}{n+2+b} & \cdots & \frac{1}{n+n+b} \end{vmatrix} = \frac{\Gamma_2(b+2)\Gamma_2(2n+b+2)}{\Gamma_2(n+b+2)^2\Gamma_2(n+1)^2}.$$

This follows from a result of Cauchy (see [8, Chap. 7, #3]) and is also noted by Gosper in [4].

(vi) $\Gamma_2(\frac{1}{2}) = A^{3/2}\pi^{1/4}e^{-1/8}2^{-1/24}$.

This is a first generalization of $\Gamma(\frac{1}{2}) = \sqrt{\pi}$; however, it lacks a geometric interpretation. We will find this interpretation by expressing $\Gamma_n(\frac{1}{2})$ in terms of $\det \Delta_m$ the determinant of the standard Laplacian of the m-sphere S^m.

Our result is Theorem 1.1.

THEOREM 1.1. *Let n be a positive integer, then there are computable rational numbers*

$$a_n, b_n, c_n, q_{n,1}, q_{n,2}, \cdots, q_{n,n} \quad \text{where } q_{n,n} = \frac{2^n - 1}{2^{n+1}}, \quad a_n \neq 0 \text{ s.t.},$$

$$\Gamma_n(\tfrac{1}{2}) = a_n^{b_n} e^{c_n} \prod_{m=1}^{n} (\det \Delta_m)^{q_{n,m}}. \tag{4}$$

In particular

$$\Gamma_2(\tfrac{1}{2}) = \det \Delta_2^{3/8} \det \Delta_1^{1/8}\, 2^{-7/24} e^{-3/16} = 1.2451432493632740351800384317993 18\cdots,$$

$$\Gamma_3(\tfrac{1}{2}) = \det \Delta_3^{7/16} \det \Delta_2^{31/64} \det \Delta_1^{-15/64} 2^{-11/48} e^{-31/128}.$$

It is seen that (4) can be written as

$$\begin{pmatrix} \log(\Gamma_1(\frac{1}{2}) a_1^{-b_1} e^{-c_1}) \\ \log(\Gamma_2(\frac{1}{2}) a_2^{-b_2} e^{-c_2}) \\ \vdots \\ \log(\Gamma_n(\frac{1}{2}) a_n^{-b_n} e^{-c_n}) \end{pmatrix} = \begin{pmatrix} \frac{1}{4} & 0 & \cdots & 0 \\ q_{2,1} & \frac{3}{8} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ q_{n,1} & q_{n,2} & \cdots & (2^n-1)/2^{n+1} \end{pmatrix} \begin{pmatrix} \log \det \Delta_1 \\ \log \det \Delta_2 \\ \vdots \\ \log \det \Delta_n \end{pmatrix}. \tag{5}$$

Thus (5) can be inverted to yield Theorem 1.2.

THEOREM 1.2. *For n be a positive integer there are computable rational numbers*

$$A_n, B_n, C_n, Q_{n,1}, Q_{n,2}, \cdots, Q_{n,n} \quad \text{where } Q_{n,n} = \frac{2^{n+1}}{2^n - 1}, \quad A_n \neq 0 \text{ s.t.},$$

$$\det \Delta_n = A_n^{B_n} e^{C_n} \prod_{m=1}^{n} \Gamma_m(\tfrac{1}{2})^{Q_{n,m}}.$$

In particular

$$\det \Delta_2 = \Gamma_2(\tfrac{1}{2})^{8/3} 2^{1/9} e^{1/2} = 3.19531148605918608395401893032062586902\cdots.$$

This has interest in light of a new result of Osgood, Phillips and Sarnak [7] (also Onofri [6]). They have shown that of all metrics on the 2-sphere of given constant area 4π, the metric of constant curvature 1 has the unique maximum value of $\det \Delta$. Theorem 1.2 computes this maximum value. As noted in [7], this result relates to recent progress in Superstring Theory in that it studies how $\det \Delta$ varies with the conformal structure.

The proof of Theorem 1.1 is in two parts. First, one expresses $\Gamma_n(\frac{1}{2})$ in terms of the Riemann ζ-function at negative integers, that is, as follows.

THEOREM 1.3. *There are rational numbers*

$$\alpha_n, \beta_n, \sigma_{n,0}, \sigma_{n,1}, \cdots, \sigma_{n,n-1} \quad \text{with } \sigma_{n,n-1} = -\frac{2^{n-1}-1}{2^{n-1}(n-1)!}$$

such that

$$\Gamma_n(\tfrac{1}{2}) = \alpha_n^{\beta_n} \prod_{m=0}^{n-1} e^{\sigma_{n,m}\zeta'(-m)}.$$

This will be proved in § 2. Second, one expresses $\det \Delta_m$ in terms of $\zeta'(s)$ at negative integers.

THEOREM 1.4. *There are rational numbers*

$$\gamma_n, \tau_{n,0}, \tau_{n,1}, \cdots, \tau_{n,n-1}, \quad \text{with } \tau_{n,n-1} = -\frac{4}{(n-1)!} \text{ s.t.},$$

$$\det \Delta_n = e^{\gamma_n} \prod_{m=0}^{n-1} e^{\tau_{n,m}\zeta'(-m)}.$$

Theorem 1.4 will be proved in § 3.

It is clear that the first part of Theorem 1.1 follows from Theorem 1.3 and Theorem 1.4. In § 4, the formulas of §§ 2 and 3 will be specialized to the cases $n = 2, 3$ and the second part of Theorem 1.1 will be proved.

2. Multiple Gamma Functions. The aim of this section is to give a proof of Theorem 1.2. The details of this, however, are quite technical. For clarity of exposition, we first prove the formula

$$\Gamma(\tfrac{1}{2}) = 2^{-1/2} e^{-\zeta'(0)}$$

using the same steps as in the theorem. The main proof will be a relatively straightforward generalization.

Remark. Since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ we will obtain a new proof of the classical result

$$\zeta'(0) = -\log \sqrt{2\pi}$$

which does not rely on other integral representations of $\zeta(s)$ (e.g. [13, p. 271]).

LEMMA 2.1. *Let* $\zeta(s, a) = \sum_{k=0}^{\infty} (k+a)^{-s}$, $0 < a < 1$, $\operatorname{Re}(s) > 1$ *be the Hurwitz* ζ*-function; then*

$$\Gamma(a) = \frac{e^{\zeta'(0,a)}}{R}, \qquad R \text{ constant}$$

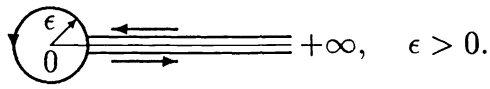
where

$$\zeta'(s, a) = \frac{\partial}{\partial s}\zeta(s, a).$$

Proof. First we note the integral formula [13]

$$\frac{i\Gamma(1-s)}{2\pi} \int_C (-z)^{s-1} e^{-qz}\, dz = q^{-s}, \qquad q > 0$$

where C is given by



Thus we have

$$\zeta(s, a) = \frac{i\Gamma(1-s)}{2\pi} \int_C \frac{e^{-az}(-z)^{s-1}}{1-e^{-z}}\, dz.$$

Therefore $\zeta(s, a)$ is analytically continued for all $s \neq 1$, $(a > 0)$. Now clearly:

$$\zeta(s, a+1) = \zeta(s, a) - a^{-s},$$

$$\zeta'(s, a+1) = \zeta(s, a) + a^{-s} \log a,$$

$$\zeta'(0, a+1) = \zeta'(0, a) + \log a.$$

Letting

$$G(a) = e^{\zeta'(0,a)}$$

we have

$$G(a+1) = aG(a)$$

and

$$\frac{d^2}{da^2} \log G(a) = \frac{d^2}{da^2} \frac{d}{ds} \zeta(s, a)\bigg|_{s=0} = \sum_{k=0}^{\infty} \frac{1}{(k+a)^2} > 0, \qquad a > 0.$$

And by the analytic continuation of $\zeta(s, a)$ one sees that $G(a)$ is C^∞ on $\mathbf{R}^+$. So, by the Bohr-Mollerup Theorem

$$G(a) = \Gamma(a)R, \qquad R \text{ constant.} \qquad \square$$

Note that $R = e^{\zeta'(0)}$ since $\zeta(s, 1) = \zeta(s)$ and so

$$R = G(1) = e^{\zeta'(0,1)}.$$

Next we prove the *Duplication Formula.*

LEMMA 2.2.

$$\Gamma(a)\Gamma(a+\tfrac{1}{2}) = 2^{(1/2)-2a}\, e^{-\zeta'(0)}\Gamma(2a).$$

Proof.

$$\zeta(s, a) + \zeta\left(s, a+\frac{1}{2}\right) = \sum_{k=0}^{\infty} \frac{2^s}{(2k+2a)^s} + \sum_{k=0}^{\infty} \frac{2^s}{(2k+1+2a)^s} = 2^s \zeta(s, 2a).$$

Thus, as in Lemma 2.1

$$G(a)G(a+\tfrac{1}{2}) = 2^{\zeta(0,2a)G(2a)}.$$

As is well known

$$\zeta(0, a) = -B_1(a)$$

where

$$B_1(a) = a - \tfrac{1}{2}$$

is the first Bernoulli polynomial, and the result follows. $\square$

On substituting $a=\frac{1}{2}$ in Lemma 2.1, we obtain

$$\Gamma(\tfrac{1}{2})=2^{-1/2}\,e^{-\zeta(0)}.$$

We now return to the proof of Theorem 1.2.

To study the Multiple Gamma Function we use the *Multiple Hurwitz Zeta Function* [2]

$$\begin{aligned}\zeta_n(s,a)&=\sum_{k_1,\cdots,k_n}(a+k_1+\cdots+k_n)^{-s}\\&=\sum_{k=0}^{\infty}\binom{k+n-1}{n-1}\Big/(k+a)^s,\qquad a>0,\quad \operatorname{Re}(s)>n.\end{aligned}$$

The first observation is Proposition 2.1.

PROPOSITION 2.1. *$\zeta_n(s,a)$ can be continued to a holomorphic function for $s\neq 1,2,\cdots,n$, $a>0$.*

Proof. By the integral formula above one has the integral representation

$$\zeta_n(s,a)=\frac{i\Gamma(1-s)}{2\pi}\int_C\frac{e^{-az}(-z)^{s-1}}{(1-e^{-z})^m}\,dz.$$

The integral is valid for $a>0$ and all s, so $\zeta(s,a)$ has possible poles only at the poles of $\Gamma(1-s)$, i.e., 1, 2, $\cdots$. But by the series definition $\zeta_n(s,a)$ is holomorphic for $\operatorname{Re}(s)>n$. □

Now define

$$G_n(a)=e^{\zeta_n'(0,a)}$$

where

$$\zeta_n'(s,a)=\frac{\partial}{\partial s}\zeta_n(s,a).$$

The basic properties of $G_n(a)$ are now given by the following proposition.

PROPOSITION 2.2. (a) $G_{n+1}(a+1)=(G_{n+1}(a))/G_n(a)$.

(b) *$G_n(a)$ can be continued to a meromorphic function on $\mathbf{C}$ with poles at the negative integers and a simple pole at zero.*

(c) *Let $R_n=\lim_{a\to 0}a\Gamma_n(a)$. Then $G_n(1)=R_n/R_{n-1}$, where $R_0=1$.*

Proof. (a) Follows from the identity

$$\zeta_{n+1}(s,a+1)=\zeta_{n+1}(s,a)-\zeta_n(s,a).$$

(b) Analytic continuation follows from the functional equation of (a). Further one has

$$G_{n+1}(a)=G_n(a)G_{n+1}(a+1),\quad\text{so}$$

$$\lim_{a\to 0}aG_{n+1}(a)=\Big(\lim_{a\to 0}aG_n(a)\Big)G_{n+1}(1). \tag{6}$$

Since $G_1(a)=\Gamma(a)/R$ by Lemma 2.1, the lemma follows by induction.

(c) This clearly follows from (6). □

PROPOSITION 2.3.

$$\Gamma_n(a)=\left[\prod_{m=1}^{n}R_{n-m+1}^{(-1)^m\binom{a}{m-1}}\right]G_n(a). \tag{7}$$

Proof. Denote by $g_n(a)$ the right side of (7). One shows that $g_n(a)$ satisfies the criteria of § 1:

(a) $g_n(1)=1$,

(b) $g_{n+1}(a+1)=g_{n+1}(a)/g_n(a)$,

(c) $g_n(a)^{-1}$ *is* $\mathbf{C}^\infty$ *on* $\mathbf{R}^+$,

(d) $(-1)^{n+1}(d^{n+1}/da^{n+1})\log g_n(a)\geqq 0$, $a>0$.

This is a straightforward verification as in Lemma 2.1. □

We now express the R_n in terms of the Riemann ζ-function. For this we write

$$\binom{x+n-1}{n-1}=\frac{(x+n-1)(x+n-2)\cdots(x+1)}{(n-1)!}=\sum_{m=0}^{n-1}S_{n,m}x^m$$

where $S_{n,m}$ are rational numbers related to *Stirling Numbers* [8].

Note. $S_{n,n-1}=1/(n-1)!$

PROPOSITION 2.4.

$$R_n=\exp\left(\sum_{r=1}^{n}\sum_{m=0}^{r-1}\zeta'(-m)\sum_{j=m}^{r-1}(-1)^{j-m}\binom{j}{m}S_{n,j}\right)$$

$$=\exp\left(\frac{\zeta'(-n+1)}{(n+1)!}+\sum_{r=0}^{n-2}\kappa_{n,r}\zeta'(-r)\right),$$

where $\kappa_{n,0},\cdots,\kappa_{n,n-2}$ *are rational numbers.*

Proof.

$$\zeta_r(s,a)=\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}\Big/(a+k)^s$$

$$=\sum_{j=0}^{r-1}S_{r,j}\sum_{k=0}^{\infty}\frac{k^j}{(a+k)^s}$$

$$=\sum_{j=0}^{r-1}S_{r,j}\sum_{k=0}^{\infty}\frac{(k+a-a)^j}{(k+a)^s}$$

$$=\sum_{j=0}^{r-1}S_{r,j}\sum_{m=0}^{j}(-a)^{j-m}\binom{j}{m}\zeta(s-m,a).$$

Now $G_n(1)=R_n/R_{n-1}$, $R_0=1$ and so

$$\log R_n=\sum_{r=1}^{n}\log G_n(1)=\sum_{r=1}^{n}\zeta_r'(0,1).$$

The result follows. □

COROLLARY 2.1.

$$\zeta_n(0,a)=-\sum_{m=0}^{n-1}\frac{B_{m+1}(a)}{m+1}\sum_{j=m}^{n-1}(-1)^{j-m}\binom{j}{m}S_{n,j},$$

where $B_k(a)$ *is the kth Bernoulli Polynomial* (*recall that* $B_k(a)\in\mathbf{Q}[a]$).

Proof. It follows from the well-known formula [13]

$$\zeta(-m,a)=-\frac{B_{m+1}(a)}{m+1}.$$

□

Note. This implies $\zeta_n(0,a)\in\mathbf{Q}[a]$.

We prove the *Generalized Duplication Formula.*

PROPOSITION 2.5.

$$\prod_{(t_1,t_2,\cdots,t_n)} \Gamma_n\left(a+\frac{t_1+\cdots+t_n}{2}\right)$$

$$=\Gamma_n(2a)2^{\zeta_n(0,2a)}\exp\left(-\frac{2^n-1}{(n-1)!}\zeta'(-n+1)+\sum_{m=0}^{n-2} f_{n,m}(a)\zeta'(-n+1)\right)$$

where $(t_1,\cdots,t_n)$ *runs over all combinations* $t_j=0$ *or* $1, j=1,\cdots,n$ *and* $f_{n,m}(a)\in \mathbf{Q}[a]$, $m=1,\cdots,n-2$.

Proof. As in Lemma 2.2, we have

$$\sum_{(t_1,\cdots,t_n)} \zeta_n\left(s, a+\frac{t_1+\cdots+t_n}{2}\right)=2^s\zeta_n(s,2a)$$

so

$$\prod_{(t_1,\cdots,t_n)} G_n\left(a+\frac{t_1+\cdots+t_n}{2}\right)=2^{\zeta_n(0,2a)}G_n(2a).$$

Substituting

$$G_n(a)=\Gamma_n(a)\prod_{m=1}^{n} R_{n-m+1}^{(-1)^{m+1}\binom{a}{m-1}},$$

one has that these equal

$$R_n^{2^n}\prod_{m=2}^{n}\prod_{(t_1,\cdots,t_n)} R_{n-m+1}^{(-1)^{m+1}\binom{a+\frac{t_1+\cdots+t_n}{2}}{m-1}}$$

$$=2^{\zeta_n(0,2a)}R_n\prod_{m=2}^{n} R_{n-m+1}^{(-1)^{m+1}\binom{2a}{m-1}}\Gamma_n(2a)$$

and the result follows on substituting the values of R_m found in Proposition 2.4. □

At this point we can finish off the proof of Theorem 1.3.

Proof of Theorem 1.3. Let $a=\frac{1}{2}$ in Proposition 2.5. This gives:

$$\prod_{(t_1,\cdots,t_n)} \Gamma_n\left(\frac{1}{2}+\frac{t_1+\cdots+t_n}{2}\right)=2^{\zeta_n(0,1)}\exp\left(-\frac{2^n-1}{(n-1)!}\zeta'(-n+1)\right)$$

$$\cdot\exp\left(-\sum_{m=0}^{n-2} f_{n,m}\left(\frac{1}{2}\right)\zeta'(-m)\right)\Gamma_n(1).$$

Now

$$\prod_{(t_1,\cdots,t_n)} \Gamma_n\left(\frac{1}{2}+\frac{t_1+\cdots+t_n}{2}\right)$$

$$=\Gamma_n\left(\frac{1}{2}\right)\Gamma_n\left(\frac{3}{2}\right)^{\binom{n}{2}}\cdots\Gamma_n\left(\left[\frac{n}{2}\right]+\frac{1}{2}\right)^{\binom{n}{[n/2]}} \tag{8}$$

$$\cdot\Gamma_n(1)^{\binom{n}{1}}\Gamma_n(2)^{\binom{n}{3}}\cdots\Gamma_n\left(\frac{n}{2}+\frac{(1+(-1)^{n+1})}{4}\right)^{\binom{n}{n-(1+(-1)^n)/2}}.$$

It follows easily from the recursion formula that $\Gamma_n(m) \in \mathbf{Q}$ for m a positive integer. So the second line of (8) is a rational number q. As for the first line it can be reduced as follows.

Let k be a positive odd integer; then

$$\Gamma_n\left(\frac{k}{2}\right) = \frac{\Gamma_n((k-2)/2)}{\Gamma_{n-1}((k-2)/2)} = \cdots = \frac{\Gamma_n(\frac{1}{2})}{\Gamma_{n-1}((k-2)/2)\Gamma_{n-1}((k-4)/2)\cdots\Gamma_{n-1}(\frac{1}{2})}.$$

So by induction there are integers

$$C_{n,1}, C_{n,2}, \cdots, C_{n,n-1}$$

such that

$$\Gamma_n\left(\frac{n}{2}\right) = \Gamma_n\left(\frac{1}{2}\right) \prod_{m=1}^{n-1} \Gamma_m\left(\frac{1}{2}\right)^{C_{n,m}}.$$

The first line of (8) is thus

$$\Gamma_n\left(\frac{1}{2}\right)\Gamma_n\left(\frac{3}{2}\right)^{\binom{n}{2}} \cdots \Gamma_n\left(\left[\frac{n}{2}\right]+\frac{1}{2}\right)^{\binom{n}{2[n/2]}}$$

$$= \Gamma_n\left(\frac{1}{2}\right)^{1+\binom{n}{2}+\binom{n}{4}+\cdots+\binom{n}{2[n/2]}} \prod_{m=1}^{n-1} \Gamma_m\left(\frac{1}{2}\right)^{d_{n,m}} \quad \text{where } d_{n,m} \in \mathbf{Z}.$$

As can easily be shown

$$1+\binom{n}{2}+\binom{n}{4}+\binom{n}{2[n/2]} = 2^{n-1}.$$

So for each n one gets

$$\Gamma_n\left(\frac{1}{2}\right) = \prod_{m=1}^{n-1} \Gamma_m\left(\frac{1}{2}\right)^{d_m/2^{n-1}} q^{2^{1-n}} 2^{\zeta_n(0,1)/2^{n-1}}$$

$$\cdot \exp\left(-\frac{2^n-1}{2^{n-1}(n-1)!}\zeta'(-n+1)\right) \prod_{m=0}^{n-2} \exp\left(\frac{f_{n,m}(\frac{1}{2})}{2^{n-1}}\zeta'(-m)\right).$$

Theorem 1.3 follows on substituting values for $\Gamma_m(\frac{1}{2})$, $m = 1, 2, \cdots, n-1$ found in the above formula. □

3. Determinants of Laplacians. We will now evaluate $\det \Delta_n$ in terms of the Riemann ζ-function.

It is well known [10] that Δ_n has eigenvalues

$$k(n+k-1) \quad \text{with multiplicity } \binom{k+n}{n} - \binom{k+n-2}{n}.$$

This gives

$$F_n(s) = \sum_{k=1}^{\infty} \left[\binom{k+n}{n} - \binom{k+n-2}{n}\right] \Big/ k^s(n+k-1)^s.$$

Now write

$$\binom{x+n}{n} - \binom{x+n-2}{n} = \sum_{d=0}^{n-1} T_{n,d} x^n.$$

Note. $T_{n,d} = 2S_{n,d} - S_{n-1,d}$ where $S_{n,d}$ was defined in § 2.

We therefore have

$$F_n(s)=\sum_{d=0}^{n-1} T_{n,d} \sum_{k=1}^{\infty} \frac{k^d}{k^s(n+k-1)^s},$$

so it is natural to define

$$H_d(s)=\sum_{k=1}^{\infty} \frac{k^d}{k^s(k+n)^s}.$$

PROPOSITION 3.1.

$$H_d'(0)=\sum_{k=1}^{n}(k-n)^d \log k-\frac{1}{2}\frac{(-n)^{d+1}}{d+1}\sum_{j=1}^{d}\frac{1}{j}+\zeta'(-d)+(-n)^d\sum_{r=0}^{d}\binom{d}{r}\zeta'(-r)\Big/(-n)^r.$$

Proof.

$$H_d'(s)=\sum_{k=1}^{\infty}\frac{-k^d(\log k+\log(k+n))}{k^s(k+n)^s}$$

$$=-\sum_{k=1}^{n}\frac{\log k}{k^{s-d}(k+n)^s}-\sum_{k=n+1}^{\infty}\frac{\log k}{k^{2s-d}}\left(\left(1+\frac{n}{k}\right)^{-s}+\left(1-\frac{n}{k}\right)^{d-s}\right). \tag{9}$$

Now

$$\left(1+\frac{n}{k}\right)^{-s}+\left(1-\frac{n}{k}\right)^{d-s}=\sum_{k=0}^{\infty} b_m(s)\left(\frac{n}{k}\right)^m, \qquad k>n$$

where

$$b_m(s)=\binom{-s}{m}+(-1)^m\binom{-s+d}{m}.$$

So letting

$$A(s)=-\sum_{k=1}^{n}\frac{\log k}{k^{s-d}(k+n)^s}$$

be the first term of (9), we get

$$A(0)=-\sum_{k=1}^{n} k^d \log k.$$

So we examine the second term. First note that for $\varepsilon>0$

$$\sum_{k=n+1}^{\infty}\frac{\log k}{k^{\sigma}}\ll\sum_{k=n+1}^{\infty}k^{-\sigma+\varepsilon}$$

$$<\int_{n+1}^{\infty}\frac{du}{u^{\sigma+\varepsilon}}=\frac{1}{(\sigma-1+\varepsilon)(n+1)^{\sigma-1+\varepsilon}}.$$

Letting

$$B(s)=-\sum_{m=d+2}^{\infty} b_m(s)n^m\sum_{k=n+1}^{\infty}\frac{\log k}{k^{2s+m-d}}$$

one has as $s\to 0$

$$B(s)\ll\sum_{m=d+2}^{\infty}\frac{|b_m(s)|n^m}{(2\sigma+m-d-1+\varepsilon)(n+1)^{2\sigma+m-d-1+\varepsilon}}$$

$$\ll\sum_{m=d+2}^{\infty}|b_m(s)|\frac{n^m}{(n+1)^m}.$$

So for $m \geqq d+1$

$$b_m(s) = \frac{(-s)(-s-1)\cdots(-s-m+1)}{m!}$$

$$+(-1)^{m!}\frac{(-s+d)(-s+d-1)\cdots(-s)\cdots(-s+d-m+1)}{m!}$$

$$= sP_m(s)$$

where

$$P_m(0) = \frac{(-1)^m}{m} + (-1)^d \Big/ m\binom{m-1}{d} \to 0 \quad \text{as } m \to \infty.$$

Therefore we write

$$H_d'(s) = A(s) + B(s) + \sum_{m=0}^{d+1} b_m(s) n^m \sum_{k=0}^{n} \frac{\log k}{k^{2s+m-d}}$$

$$+ \sum_{m=0}^{d+1} b_m(s) n^m \zeta'(2s+m-d).$$

Letting

$$C(s) = \sum_{m=0}^{d+1} b_m(s) n^m \sum_{k=0}^{n} \frac{\log k}{k^{2s+m-d}},$$

we have

$$C(0) = \sum_{k=0}^{n} \log k(k-n)^d + \sum_{k=0}^{n} k^d \log k$$

and note that the second term cancels with $A(0)$. Turning to

$$D(s) = \sum_{m=0}^{d+1} b_m(s) n^m \zeta'(2s+m-d),$$

we distinguish 3 cases: $m=0$, $0 < m \leqq d$ and $m = d+1$.

Now

$$b_0(s) = \binom{-s}{0} + (-1)^0 \binom{-s+d}{0} \quad \text{so } b_0(0) = 2$$

and

$$b_m(0) = (-1)^m \binom{d}{m}, \qquad 0 < m \leqq d.$$

Thus

$$D(0) = 2\zeta'(-d) + \sum_{m=1}^{d} (-n)^m \binom{d}{m} \zeta'(m-d) + b_{d+1}(s) n^{d+1} \zeta'(2s+1)\big|_{s=0}.$$

Recall that

$$\zeta(s) = \frac{1}{s-1} + a_0 + a_1(s-1)\cdots,$$

so

$$\zeta'(2s+1)=\frac{-1}{4s^2}+2a_1s+\cdots.$$

While

$$b_{d+1}(s)=(-1)^{d+1}\frac{s(s+1)\cdots(s+d)}{(d+1)!}+\frac{s(s-1)\cdots(s-d)}{(d+1)!}$$
$$=\left[\frac{2(-1)^{d+1}}{d+1}\sum_{j=1}^{d}\frac{1}{j}\right]s^2+\mathcal{O}(s^3)\quad\text{as } s\to 0,$$

we have

$$\lim_{s\to 0} b_{d+1}n^{d+1}\zeta'(2s+1)=\frac{n^{d+1}}{2(d+1)}(-1)^d\sum_{j=1}^{d}\frac{1}{j}.$$ □

We can now prove Theorem 1.4, i.e., that there are rational numbers

$$\gamma_n,\ \tau_{n,0},\ \tau_{n,1},\ \cdots,\ \tau_{n,n-1},\quad\text{with } \tau_{n,n-1}=\frac{-4}{(n-1)!}\text{ s.t.,}$$

$$\det\Delta_n=e^{\gamma_n}\prod_{m=0}^{n-1}e^{\tau_{n,m}\zeta'(-m)}.$$

Proof. By the proposition

$$F_n'(0)=\sum_{d=0}^{n-1}T_{n,d}\sum_{k=1}^{n-1}(k-n+1)^d\log k-\frac{1}{2}\sum_{d=0}^{n-1}T_{n,d}\frac{(1-n)^{d+1}}{d+1}\sum_{j=1}^{d}\frac{1}{j}$$
$$+\sum_{r=0}^{n-1}\zeta'(-r)\left(T_{n,r}+\sum_{d=r}^{n-1}T_{n,r}(1-n)^{d-r}\binom{n-1}{r}\right).$$

Now $T_{n,d}=2S_{n,d}-S_{n-1,d}$; so

$$\sum_{d=0}^{n-1}(2S_{n,d}-S_{n-1,d})\sum_{k=1}^{n-1}(k-n+1)^d\log k$$
$$=\sum_{k=1}^{n-2}\log k\sum_{d=0}^{n-1}[2S_{n,d}(k-n+1)^d-S_{n-1,d}(k-n+1)^d]$$
$$=\sum_{k=1}^{n-2}\log k\left[2\binom{k-n+1+n-1}{n-1}-\binom{k-n+1+n-2}{n-2}\right]=0.$$

Furthermore

$$T_{n,n-1}=2S_{n,n-1}=\frac{2}{(n-1)!},$$

so

$$\sum_{r=0}^{n-1}\zeta'(-r)\left[T_{n,r}+\sum_{d=r}^{n-1}T_{n,r}(1-n)^{d-r}\binom{n-1}{r}\right]$$
$$=\frac{4}{(n-1)!}\zeta'(1-n)-\sum_{m=0}^{n-2}\tau_{n,m}\zeta'(-m),\ \tau_{n,m}\in\mathbf{Q}.$$

Letting

$$\tau_n = \frac{1}{2}\sum_{d=0}^{n-1} T_{n,d}\frac{(1-n)^{d+1}}{d+1}\sum_{j=1}^{d}\frac{1}{j} \in \mathbf{Q}$$

yields

$$\det \Delta_n = e^{-F_n'(0)} = \exp\left(\frac{-4}{(n-1)!}\zeta'(1-n) + \sum_{m=0}^{n-2}\tau_{n,m}\zeta'(-m)\right)e^{\tau_n}. \qquad \square$$

4. The cases $n=2, 3$. In this section we specialize the results of §§ 2 and 3 to the cases $n=2, 3$. Since these follow by substitution we omit the proofs.

PROPOSITION 4.1.

$$R_1 = e^{\zeta'(0)}, \qquad R_2 = e^{\zeta'(0)+\zeta(-1)}, \qquad R_3 = \exp\left(\zeta'(0)+\tfrac{3}{2}\zeta'(-1)+\tfrac{1}{2}\zeta'(-2)\right),$$

$$\zeta_1(0, a) = \tfrac{1}{2} - a, \qquad \zeta_2(0, a) = (1-a)\zeta(0, a) + \zeta(-1, a),$$

$$\zeta_3(0, a) = \frac{1}{2}\zeta(-2, a) + \left(\frac{3}{2} - a\right)\zeta(-1, a) + \left(\frac{a^2}{2} - \frac{3a}{2} + 1\right)\zeta(0, a)$$

and

$$\zeta(0, a) = \frac{1}{2} - a, \qquad \zeta(-1, a) = -\frac{a^2}{2} + \frac{a}{2} - \frac{1}{12},$$

$$\zeta(-2, a) = -\frac{a^3}{3} + \frac{a^2}{2} - \frac{a}{6}.$$

So

$$\zeta_1(0, 1) = \frac{-1}{12}, \qquad \zeta_2(0, 1) = \frac{-1}{12}, \qquad \zeta_3(0, 1) = \frac{-1}{24}.$$

The duplication formulas for $n=2, 3$ are given by the following proposition.

PROPOSITION 4.2.

$$\Gamma_2(a)\Gamma_2(a+\tfrac{1}{2})^2\Gamma_2(a+1) = 2^{\zeta_2(0,2a)}\exp\left(-3\zeta'(-1) + (2a-1)\zeta'(0)\right)\Gamma_2(2a),$$

$$\Gamma_3(a)\Gamma_3(a+\tfrac{1}{2})^3\Gamma_3(a+1)^3\Gamma_3(a+\tfrac{3}{2})$$

$$= 2^{\zeta_3(0,2a)}\exp\left(-\tfrac{7}{2}\zeta'(-2) + (6a-\tfrac{9}{2})\zeta'(-1) + (-2a^2+3a-1)\zeta'(0)\right)\Gamma_3(2a).$$

And so one gets Proposition 4.3.

PROPOSITION 4.3.

$$\Gamma_2(\tfrac{1}{2}) = 2^{-7/24}\exp\left(-\tfrac{3}{2}\zeta'(-1) - \tfrac{1}{2}\zeta'(0)\right),$$

$$\Gamma_3(\tfrac{1}{2}) = 2^{-11/48}\exp\left(-\tfrac{7}{8}\zeta'(-2) - \tfrac{3}{2}\zeta'(-1) - \tfrac{3}{8}\zeta'(0)\right).$$

On the other hand, one has the following proposition.

PROPOSITION 4.4.

$$F_1'(0) = 4\zeta'(0),$$

$$F_2'(0) = 4\zeta'(-1) - \tfrac{1}{2},$$

$$F_3'(0) = 2\zeta'(-2) - \zeta'(-1) + 3\zeta'(0).$$

So one gets Proposition 4.5.

PROPOSITION 4.5.

$$e^{\zeta'(0)} = (\det \Delta_1)^{-1/4},$$
$$e^{\zeta'(-1)} = (\det \Delta_2)^{-1/4}\, e^{1/8},$$
$$e^{\zeta'(-2)} = (\det \Delta_3)^{-1/2}(\det \Delta_2)^{-1/8}(\det \Delta_1)^{3/8}\, e^{1/16}.$$

The second part of Theorem 1.1 now follows from substituting Proposition 4.5 in Proposition 4.3.

This completes the proof of the main results of the paper. We conclude by proving the formula for Kinkelin's constant given in § 1.

PROPOSITION 4.6.

$$A = e^{1/12-\zeta'(-1)}.$$

Proof. Recall the Hurwitz ζ-function

$$\zeta(s, a) = \sum_{n=0}^{\infty} (n+a)^{-s}, \qquad s>1;$$

then

$$\sum_{k=1}^{n-1} k^s = \zeta(-s) - \zeta(-s, n).$$

Thus

$$\sum_{k=1}^{n-1} k \log k = -\zeta'(-1) + \zeta'(-s, n). \tag{10}$$

We have the representation

$$\zeta(s, a) = \frac{a^{1-s}}{s-1} + \frac{1}{2a^s} + \frac{sB_2}{2a^{s+1}} - \frac{s(s+1)(s+2)}{6} \int_0^{\infty} \frac{B_3(u-[u])}{(u+a)^{s+3}}\, du, \qquad \operatorname{Re}(s) > -2. \tag{11}$$

This is easily shown by expanding $B_3(u-[u])$ and collecting terms. This formula is an extension of the well-known representation for the Riemann ζ-function

$$\zeta(s) = \frac{s}{s-1} - s \int_0^{\infty} \frac{u-[u]}{u^{s+1}}\, du.$$

Differentiating (11) gives

$$\zeta'(s, a) = -\frac{a^{-s} \log a}{2} - \frac{a^{1-s}}{(1-s)^2} + \frac{a^{1-s} \log a}{1-s} + \frac{B_2 a^{-s-1}}{2} - \frac{B_2 a^{-s-1} \log a}{2} + \mathcal{O}\left(\frac{1}{a}\right)$$

and so

$$\zeta'(-1, n) = -\frac{n \log n}{2} + \frac{n^2 \log n}{2} + \frac{\log n}{12} - \frac{n^2}{4} + \frac{1}{12} + \mathcal{O}\left(\frac{1}{n}\right).$$

The result follows on substituting this in (10). □

Acknowledgments. I would like to thank Peter Sarnak for suggesting this problem to me and for his help on § 3, and Bill Gosper for his help with MACSYMA.

REFERENCES

[1] E. ARTIN, *The Gamma Function*, Holt, Rinehart and Winston, New York, 1964.

[2] E. W. BARNES, *On the theory of the Multiple Gamma Function*, Trans. Cambridge Philos. Soc., 19 (1904), pp. 374–425.

[3] E. W. BARNES, *The theory of the G-function*, Quart. J. Math., 31 (1899), pp. 264–314.
[4] R. W. GOSPER, *Higher factorials*, unpublished notes.
[5] KINKELIN, *Ueber eine mit der Gamma Function verwandte Transcendente und deren Anwendung auf die Integralrechnung*, J. Reine Angew. Math., 57 (1860), pp. 122–158.
[6] E. ONOFRI, *On the positivity of the effective action in a theory of random surface*, Comm. Math. Phys., 86 (1982), pp. 321–326.
[7] B. OSGOOD, R. PHILLIPS AND P. SARNAK, *Extremals of determinants of Laplacians*, preprint 1986.
[8] G. POLYA AND G. SZEGO, *Problems and Theorems in Analysis* I *and* II, Springer-Verlag, New York, 1972.
[9] P. SARNAK, *Determinants of Laplacians*, preprint, 1986.
[10] T. SHINTANI, *A proof of the classical Kronecker Limit Formula*, Tokyo J. Math., 3 (1980), pp. 191–199.
[11] A. TERRAS, *Harmonic Analysis on Symmetric Spaces and Applications* I, Springer-Verlag, New York, 1985.
[12] M. F. VIGNÉRAS, *L'équation fonctionelle de la fonction zéta de Selberg du groupe modulaire* SL(2, **Z**), Astérisque, 61 (1979), pp. 235–249.
[13] E. T. WHITTAKER AND G. N. WATSON, *A Course of Modern Analysis*, Cambridge University Press, London, 1965.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

COMBUSTION IN A POROUS MEDIUM*

AVNER FRIEDMAN† AND ATHANASSIOS E. TZAVARAS‡

Abstract. A nonlinear time-dependent system of partial differential equations describing combustion in a porous medium is studied. Existence of a solution is established and its asymptotic behavior, as $t \to \infty$, is obtained.

Key words. combustion, porous medium, reaction rate

AMS(MOS) subject classifications. 80A20, 80A25, 35K60, 35K65

1. The physical problem. Porous medium combustion occurs in a number of situations including the burning of coal [7], the burning of cigarettes [1], the use of catalytic converters as exhaust filters [8] and the smouldering of polyurethane [6].

A model for combustion in a porous medium was developed by Lawson and Norbury [3], [4] and, more recently, by Norbury and Stuart [5]. Norbury and Stuart developed a three-dimensional model in a situation in which the chemical process is

$$\text{solid} + O_2 \to \text{heat} + CO_2 + \text{ash}.$$

Their model represents conservation of mass and energy for both the gas and solid species, while the fluid flow is governed by Darcy's law and the ideal gas law. Subsequently, in the case of one-space-dimensional combustion, they used a number of asymptotic considerations, including, most notably, large activation energy asymptotics (of Frank-Kamenetskii [2]) to arrive at a simplified model of the form

$$(1.1)\quad \frac{\partial \sigma}{\partial t} = -\lambda r, \quad \mu \frac{\partial w}{\partial x} = u - w, \quad \sigma \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left((1+du^3)\frac{\partial u}{\partial x}\right) + w - u + r, \quad \frac{\partial g}{\partial x} = -\frac{a}{\mu} r,$$

where λ, μ, d, and a are positive constants, r is the reaction rate

$$(1.2)\qquad r = H(\sigma - \tau)H(u - u_c)H(g)\mu^{1/2} g f(w),$$

H is the Heaviside function,

$$f(w) = w^{\nu} \quad \text{where typically } \nu \sim 2.0$$

and τ, u_c are positive constants. Finally, the boundary conditions are

$$(1.3)\qquad u(\pm N, t) = u_a, \quad w(-N, t) = u_a, \quad g(-N, t) = g_a$$

where u_a, g_a are positive constants and N is a positive number or $+\infty$, and the initial conditions are

$$(1.4)\qquad \sigma(x, 0) = \sigma_0(x), \qquad u(x, 0) = u_0(x)$$

where $\sigma_0(x) > 0$, $u_0(x) > 0$.

In the above equations, up to scaling, u represents the temperature of the solid, w represents the temperature of the gas, σ is the heat capacity of the solid, and $g = u\alpha$ where α is the mass of the oxygen per unit volume. An interesting feature of the model, the mathematical ramifications of which we explore here, is the presence of a mechanism that switches off the reaction when either the temperature, or the heat capacity of the solid, or the mass of oxygen per unit volume drops below a certain level.

* Received by the editors March 9, 1987; accepted for publication May 20, 1987. This work was partially supported by National Science Foundation grants DMS-8420896 and DMS-8501397.

† Department of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

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

In this paper we shall establish the existence of a solution of (1.1)-(1.4) for all $t>0$; we shall also prove that the reaction rate r vanishes identically if t is sufficiently large. The occurrence of σ as a coefficient of $\partial u/\partial t$ in the parabolic equation for u in (1.1) turns out to be quite favorable for the proofs.

In §§ 2-5 we consider the case $N<\infty$ and in § 6 we consider the case $N=\infty$. In § 2 we formulate the mathematical problem more generally, and derive a priori lower bounds on u, w. In § 3 we derive various integral and pointwise bounds on the solution, assuming its existence. In § 4 we establish the existence of a solution, using the a priori estimates of §§ 2 and 3, and in § 5 we study the asymptotic behavior of the solution as $t\to\infty$.

2. Mathematical formulation; initial estimates. Let N be any positive number and let

$$\Omega=\{-N<x<N\},\qquad \Omega_t=\{(x,s),\ x\in\Omega,\ 0<s<t\}.$$

In studying (1.1)-(1.4) we assume for simplicity that $\lambda=1$, $\mu=1$, $a=1$ and $\tau=1$ (the proofs remain the same for general λ, μ, a, τ). We also replace $1+du^3$ by a general function $k(u)$ positive-valued for $u\geqq 0$ and take $f(w)$ to be a general positive-valued function for $w>0$. Thus the system has the form

$$\sigma_t=-r, \tag{2.1}$$

$$w_x=u-w, \tag{2.2}$$

$$\sigma u_t=(k(u)u_x)_x+w-u+r, \tag{2.3}$$

$$g_x=-r \tag{2.4}$$

where

$$r=r(x,t)=H(\sigma-1)H(u-u_c)H(g)gf(w) \tag{2.5}$$

with the initial conditions

$$\sigma(x,0)=\sigma_0(x), \tag{2.6}$$

$$u(x,0)=u_0(x) \tag{2.7}$$

and the boundary conditions

$$w(-N,t)=u_a, \tag{2.8}$$

$$u(\pm N,t)=u_a, \tag{2.9}$$

$$g(-N,t)=g_a. \tag{2.10}$$

Here

$$H(\xi)=0 \quad \text{if } \xi<0,\qquad H(\xi)=1 \quad \text{if } \xi>0.$$

We make the following assumptions:

$$g_a>0,\quad u_a>0,\quad u_0(\pm N)=u_a,$$

$$\sigma_0(x)>0,\quad u_0(x)>0 \quad \text{if } -N\leqq x\leqq N, \tag{2.11}$$

σ_0 is Holder continuous in $\{-N\leqq x\leqq N\}$, $\qquad u_0$ belongs to $W^{2,\infty}(-N,N)$,

$$k(u)\in C^{1+\alpha}(\mathbb{R}),\quad k(u)\geqq k_0>0 \quad (\text{for some } \alpha\in(0,1)), \tag{2.12}$$

$$f(w)\in C^{0,1}(\mathbb{R}),\qquad f(w)>0. \tag{2.13}$$

The function $H(u-u_c)$ is understood as a selection from the graph $H(u-u_c)$, and $H(\sigma-1)=0$ if $\sigma-1\leqq 0$, $H(g)=0$ if $g\leqq 0$. Since $\sigma(x,t)$ is decreasing in t, also $H(\sigma(x,t)-1)$ is decreasing in t. Similarly $H(g(x,t))$ is decreasing in x, and $H(g)g=g^+$.

In this section and in the following one we assume that a solution exists for all $0<t<T$ (for some $T>0$) and derive a priori estimates. These estimates will be used in § 4 to establish the existence of a solution. From (2.1), (2.6) we get

$$\begin{aligned}\sigma(x,t)&=1+\left[\sigma_0(x)-1-\int_0^t r(x,s)\,ds\right]^+ \quad \text{if } \sigma_0(x)>1\\ &=\sigma_0(x) \quad \text{if } \sigma(x)\leqq 1.\end{aligned}\tag{2.14}$$

Let

(2.15)

$$t_0(x)=\begin{cases}0 & \text{if } \sigma_0(x)\leqq 1,\\ \sup\left\{t;\int_0^t H(u(x,s)-u_c)g^+(x,s)f(w(x,s))\,ds\leqq\sigma_0(x)-1\right\} & \text{if } \sigma_0(x)>1.\end{cases}$$

From the choice of $H(\sigma-1)$ mentioned above we have

$$H(\sigma-1)H(u-u_c)g^+f(w)=\chi\{t\leqq t_0(x)\}H(u-u_c)g^+f(w)\equiv r.\tag{2.16}$$

Therefore

$$\int_0^t r(x,s)\,ds=\int_0^{t\wedge t_0(x)} H(u-u_c)g^+f(w)\,ds\leqq\sigma_0(x)\leqq C\tag{2.17}$$

where C is a positive constant independent of (x,t), and $t\wedge s=\min\{t,s\}$.

LEMMA 2.1. *There holds*

$$w(x,t)\geqq u_a\,e^{-2N},\tag{2.18}$$

$$u(x,t)\geqq u_a\,e^{-2N}\tag{2.19}$$

in Ω_T.

Proof. For some δ small enough $u\geqq 0$ in Ω_δ. From (2.2), (2.8) we then find that

$$w(x,t)\geqq u_a\,e^{-(x+N)}.\tag{2.20}$$

Then also

$$Lu\equiv\sigma(x,t)u_t-(k(u)u_x)_x+u=w+r\geqq u_a\,e^{-2N}\tag{2.21}$$

in Ω_δ. The constant $z\equiv u_a\,e^{-2N}$ satisfies

$$Lz=z\leqq Lu \quad \text{in } \Omega_\delta$$

and $z\leqq u$ on the parabolic boundary of Ω_δ. By comparison, it follows that

$$u\geqq z=u_a\,e^{-2N}\tag{2.22}$$

in Ω_δ. We can now continue to increase δ step-by-step and to establish (2.20), (2.22) in the corresponding domains Ω_δ; this completes the proof of the lemma.

3. A priori estimates. Set

$$\tilde{w} = w - u_a, \qquad \tilde{u} = u = u_a.$$

LEMMA 3.1. *There holds*

$$\begin{aligned} &\frac{1}{2}\int_{\Omega\times\{t\}} \sigma\tilde{u}^2\,dx + \frac{1}{4}\int\int_{\Omega_t} r\tilde{u}^2 + \int\int_{\Omega_t} k(u)u_x^2 \\ &\quad + \frac{1}{2}\int_0^t (\tilde{w}(N,t))^2\,dt + \int\int_{\Omega_t} (\tilde{w}-\tilde{u})^2 \leqq C \qquad (0<t<T) \end{aligned} \tag{3.1}$$

where C is a positive constant independent of T.

Proof. Multiplying (2.3) by $\tilde{u}$ and integrating over Ω_t, we get

$$\frac{1}{2}\int\int_{\Omega_t} \sigma(\tilde{u}^2)_t + \int\int_{\Omega_t} k(u)(\tilde{u}_x)^2 = \int\int_{\Omega_t} (\tilde{w}-\tilde{u})\tilde{u} + \int\int_{\Omega_t} r\tilde{u}. \tag{3.2}$$

The first term on the left-hand side is equal to

$$\frac{1}{2}\int_{\Omega\times\{t\}} \sigma(\tilde{u}^2) - \frac{1}{2}\int_{\Omega\times\{0\}} \sigma\tilde{u}^2 + \frac{1}{2}\int\int_{\Omega_t} r\tilde{u}^2 \qquad (\text{since } \sigma_t = -r).$$

The last term on the right-hand side of (3.2) is bounded above by

$$\int\int_{\Omega_t} r\left(\frac{1}{4}\tilde{u}^2 + 1\right) \leqq \frac{1}{4}\int\int_{\Omega_t} r\tilde{u}^2 + \int_\Omega \left(\int_0^t r\,dt\right) dx \leqq \frac{1}{4}\int\int_{\Omega_t} r\tilde{u}^2 + C,$$

by (2.17). We therefore obtain from (3.2)

$$\frac{1}{2}\int_{\Omega\times\{t\}} \sigma\tilde{u}^2\,dx + \frac{1}{4}\int\int_{\Omega_t} r\tilde{u}^2 + \int\int_{\Omega_t} k(u)u_x^2 \leqq \int\int_{\Omega_t} (\tilde{w}-\tilde{u})\tilde{u} + C \tag{3.3}$$

where C is a constant independent of T.

Next multiply (2.2) by $\tilde{w}$ and integrate over Ω_t

$$\frac{1}{2}\int\int_{\Omega_t} (\tilde{w}^2)_x = \int\int_{\Omega_t} (\tilde{u}-\tilde{w})\tilde{w}, \tag{3.4}$$

and the left-hand side is equal to

$$\frac{1}{2}\int_0^t \tilde{w}^2(N,t)\,dt.$$

When we add (3.4) to (3.3), the inequality (3.1) follows.

LEMMA 3.2. *There holds*

$$w \leqq C, \tag{3.5}$$

$$r \leqq C, \tag{3.6}$$

$$u \leqq C \tag{3.7}$$

in Ω_T, where C is a constant independent of T.

Proof. From Lemma 3.1 and (2.14),

$$\int_\Omega \tilde{u}^2(x,t)\,dx \leqq C. \tag{3.8}$$

Hence, solving for w from (2.2), (2.8) we easily obtain the estimate (3.5).

From (2.4), (2.10) we conclude that

$$0 \leqq g \leqq g_a. \tag{3.9}$$

When we combine this with (3.5), the estimate (3.6) follows.

From (3.5), (3.6) we deduce that

$$Lu \leqq 2C$$

where L is the parabolic operator defined in (2.21). If z is a constant larger than $2C$ and larger than u on the parabolic boundary of Ω_T then, since $Lz > 2C$, we get by comparison that $u < z$ in Ω_T. This yields the estimate (3.7).

LEMMA 3.3. *There holds*

$$\iint_{\Omega_T} u_t^2 + \int_{\Omega} u_x^2(x, t)\, dx \leqq C \quad \text{for } 0 < t < T \tag{3.10}$$

where C is a constant independent of T.

Proof. Multiplying (2.3) by $k(u)u_t$ and integrating over Ω_t, we get

$$\int_{\Omega_t} \sigma k(u) u_t^2 + \frac{1}{2} \int_{\Omega \times \{t\}} k^2(u) u_x^2\, dx = \iint_{\Omega_t} (\tilde{w} - \tilde{u} + r) k(u)_t + \frac{1}{2} \int_{\Omega} k^2(u_0) u_{0,x}^2\, dx. \tag{3.11}$$

Since $k(u)$ is bounded (by Lemma 3.2), the first integrate on the right-hand side of (3.11) can be estimated by

$$\left| \iint_{\Omega_t} (\tilde{w} - \tilde{u} + r) k(u) u_t \right| \leqq \frac{1}{2} \iint_{\Omega_t} \sigma k(u) u_t^2 + C \iint_{\Omega_t} (\tilde{w} - \tilde{u} + r)^2. \tag{3.12}$$

When we use the estimates

$$\iint_{\Omega_t} (\tilde{w} - \tilde{u})^2 \leqq C \qquad \text{(by Lemma 3.1)},$$

$$\iint_{\Omega_t} r^2 \leqq C \iint_{\Omega_t} r \leqq C \quad \text{(by (3.6), (2.17))}$$

on the right-hand side of (3.12) and then substitute (3.12) into (3.11), the assertion (3.10) follows.

From Lemma 3.3 and standard interpolation we get the following lemma.

LEMMA 3.4. *There exists $0 < \alpha < 1$ and $C > 0$ independent of T such that*

$$|u|_{C^{\alpha}(\Omega_T)} \leqq C. \tag{3.13}$$

Set

$$K(u) = \int_0^u k(v)\, dv. \tag{3.14}$$

Then

$$\frac{\sigma}{k(u)} K(u)_t = K(u)_{xx} + h \tag{3.15}$$

where $h = w - u + r$ satisfies

$$|h|_{L^{\infty}(\Omega_T)} \leqq C, \quad C \text{ independent of } T.$$

Applying L^p-theory we get the following lemma.

LEMMA 3.5. *For any $p>1$ there exists a constant C depending on p, T such that*

$$|u_x|_{L^p(\Omega_T)}+|u_t|_{L^p(\Omega_T)}+|u_{xx}|_{L^p(\Omega_T)}\leqq C. \tag{3.16}$$

4. Global existence. For any $\varepsilon>0$ set

$$H_\varepsilon(\xi)=\begin{cases}1 & \text{if } \xi\geqq\varepsilon,\\ \xi/\varepsilon & \text{if } 0<\xi<\varepsilon,\\ 0 & \text{if } \xi\leqq 0.\end{cases} \tag{4.1}$$

DEFINITION 4.1. We shall refer to the system (2.1)-(2.10) as problem (P); if we replace (in (2.1)-(2.5)) $H(\sigma-1)$, $H(u-u_c)$, $H(g)$ by $H_\varepsilon(\sigma-1)$, $H_\varepsilon(u-u_c)$, $H_\varepsilon(g)$, then we refer to the new system as problem (P_ε).

Remark 4.1. All the estimates derived in §§ 2 and 3 for solutions of problem (P) are valid, with the same proofs and the same constants, also for solutions of problem (P_ε).

LEMMA 4.1. *For any $T>0$, $\varepsilon>0$ there exists at most one solution of problem (P_ε) in Ω_T with $\partial u/\partial t$, $\partial u/\partial x$ in $L^\infty(\Omega_T)$.*

Proof. Suppose $(\sigma_i, w_i, u_i, g_i)$ are solutions of problem (P_ε) for $i=1,2$. Since $H_\varepsilon(\xi)$ is Lipschitz continuous, the function h in (3.15) is actually Holder continuous (for the case of problem (P_ε)) and the solutions $(\sigma_i, w_i, u_i, g_i)$ satisfy the differential equations and the initial-boundary conditions in the classical sense.

Take the difference of the equations (2.1) for σ_1 and σ_2, multiply by $\sigma_1-\sigma_2$ and integrate over Ω_t. We then easily obtain, after applying the Cauchy-Schwarz inequality,

$$\int_{\Omega\times\{t\}}(\sigma_1-\sigma_2)^2\leqq C\int\int_{\Omega_t}[(u_1-u_2)^2+(w_1-w_2)^2+(g_1-g_2)^2]. \tag{4.2}$$

Similarly we get from the equations for w_i,

$$\int\int_{\Omega_t}(w_1-w_2)^2\leqq C\int\int_{\Omega_t}(u_1-u_2)^2. \tag{4.3}$$

Next, using (2.3), we obtain for u_1-u_2,

$$\begin{aligned}\frac{1}{2}\int_{\Omega\times\{t\}}&\sigma_1(u_1-u_2)^2+\int\int_{\Omega_t}k(u_1)(u_{1,x}-u_{2,x})^2\\ &\leqq\int\int_{\Omega_t}|\sigma_1-\sigma_2|\left|\frac{\partial u_2}{\partial t}\right||u_1-u_2|+\int\int_{\Omega_t}|k(u_1)-k(u_2)||u_{2,x}||u_{1,x}-u_{2,x}|\\ &\quad+C\int\int_{\Omega_t}[(u_1-u_2)^2+(w_1-w_2)^2+(g_1-g_2)^2+(\sigma_1-\sigma_2)^2]\end{aligned}$$

whence, because of (2.12), Schwarz's inequality and the boundedness of $\partial u_2/\partial t$ and $\partial u_2/\partial x$,

$$\frac{1}{2}\int_{\Omega\times\{t\}}\sigma_1(u_1-u_2)^2\leqq C\int\int_{\Omega_t}[(u_1-u_2)^2+(w_1-w_2)^2+(g_1-g_2)^2+(\sigma_1-\sigma_2)^2]. \tag{4.4}$$

Finally, from (2.4),

$$\int\int_{\Omega_t}(g_1-g_2)^2\leqq C\int\int_{\Omega_t}[(u_1-u_2)^2+(w_1-w_2)^2+(\sigma_1-\sigma_2)^2]. \tag{4.5}$$

We now substitute $\iint (g_1-g_2)^2$ from (4.5) into (4.2), (4.4) and substitute $\iint (w_1-w_2)^2$ from (4.3) into (4.2), (4.4). Then, combining (4.2) and (4.4), we obtain the inequality

$$\int_{\Omega\times\{t\}} [(\sigma_1-\sigma_2)^2+(u_1-u_2)^2] \leqq C \int\int_{\Omega_t} [(\sigma_1-\sigma_2)^2+(u_1-u_2)^2].$$

This immediately implies that $\sigma_1-\sigma_2\equiv 0$, $u_1-u_2\equiv 0$ and then also $w_1-w_2\equiv 0$, $g_1-g_2\equiv 0$.

We shall now prove one of the main results of this paper.

THEOREM 4.2. *There exists a global solution of* (2.1)-(2.10) *satisfying* (3.16) *for any* $1<p<\infty$, $T>0$.

Proof. We first consider problem (P_ε) and establish the existence of a global solution. Choose a small $\delta>0$ and introduce the set

$$K=\{(\sigma, w, u, g)\in C(\bar{\Omega}_\delta),\ \|\sigma\|\leqq 1+\sup \sigma_0(x),\ \|w\|\leqq C_1,\ \|u\|\leqq 1+u_a + \sup u_0(x),\ \|g\|\leqq C_2\}$$

where $\|\cdot\|$ means the sup-norm in $\bar{\Omega}_\delta$ and C_1, C_2 are to be determined later on. Given any (σ, w, u, g) in K we define its image $(\bar{\sigma}, \bar{w}, \bar{u}, \bar{g})=W(\sigma, w, u, g)$ by solving

$$\bar{\sigma}_t=-H_\varepsilon(\sigma-1)H_\varepsilon(u-u_c)H_\varepsilon(g)gf(w),$$

$$\bar{w}_x=u-\bar{w},$$

$$\sigma\bar{u}_t=\frac{\partial}{\partial x}(k(u)\bar{u}_x)+\bar{w}-\bar{u}+H_\varepsilon(\sigma-1)H_\varepsilon(u-u_c)H_\varepsilon(g)gf(w),$$

$$\bar{g}_x=-H_\varepsilon(\sigma-1)H_\varepsilon(u-u_c)H_\varepsilon(g)gf(w)$$

in Ω_δ subject to the initial and boundary conditions (2.6)-(2.10).

For appropriately large C_1 and C_2 (C_2 depends on C_1) we can easily estimate

$$\|\bar{\sigma}\|\leqq 1+\sup \sigma_0(x),\quad \|\bar{w}\|\leqq C_1,\quad \|\bar{g}\|\leqq C_2.$$

While for C_1, C_2 as above and δ sufficiently small,

$$\|\bar{u}\|\leqq 1+u_a+\sup u_0(x).$$

By parabolic estimates we also have, for some $\alpha\in(0,1)$

$$|\bar{u}|_{C^\alpha(\bar{\Omega}_\delta)}\leqq C;$$

here C is a general constant which may depend on δ. Since $\sigma_0(x)$ is Hölder continuous, we then also deduce that

$$|\bar{\sigma}|_{C^\alpha(\bar{\Omega}_\delta)}\leqq C$$

with another α and C_1 and then, finally, also

$$|\bar{g}|_{C^\alpha(\bar{\Omega}_\delta)}\leqq C.$$

Hence W maps K into a compact subset. Since $(\bar{\sigma}, \bar{w}, \bar{u}, \bar{g})$ is uniquely determined by (σ, w, u, g), we can easily verify that W is a continuous mapping. Appealing to the Schauder fixed-point theorem we conclude that W has a fixed point. Thus problem (P_ε) has a solution in Ω_δ. We note that the size of δ depends only on the upper bounds on

$$|u_0|_{C^\beta(\Omega)},\qquad |\sigma_0|_{C^\beta(\Omega)} \tag{4.6}$$

for some fixed $\beta\in(0,1)$.

We now wish to extend the solution to $\delta < t < 2\delta$. We choose any initial time $t = \delta - \eta$ (η arbitrarily small) and repeat the previous step. We obtain a new solution for $\delta - \eta < t < (\delta - \eta) + \delta_1$ for some $\delta_1 > 0$, δ_1 independent of η. Since the solution constructed in Ω_δ is a classical solution with $\partial u/\partial t$, $\partial u/\partial x$ bounded (or even Holder continuous) in $\Omega_\delta \backslash \Omega_{\delta-\eta}$ (since $H_\varepsilon(\xi)$ is Lipschitz continuous) we can apply the uniqueness result of Lemma 4.1 in $\Omega_\delta \backslash \Omega_{\delta-\eta}$ and thus deduce that the new solution is an extension to $\Omega_{\delta-\eta+\delta_1}$ of the solution constructed in the first step in Ω_δ. Thus we obtain a solution of problem (P_ε) in Ω_{δ_2} where δ_2 in any number $< \delta + \delta_1$.

We can proceed in this manner step-by-step, and it only remains to show that in these steps the lengths of the t-intervals δ_i do not shrink to zero. For this it suffices to show that the quantities in (4.6) corresponding to any initial time $t = \tau$ remain uniformly bounded if $\tau \leqq T$, where T is any given positive number. But this clearly follows from Remark 4.1.

Having constructed a global solution to problem (P_ε), we denote it by $(\sigma_\varepsilon, w_\varepsilon, u_\varepsilon, g_\varepsilon)$. (Notice that we have not proved boundedness of $\partial u_\varepsilon/\partial t$, $\partial u_\varepsilon/\partial x$ at $t = 0$ and thus we cannot apply Lemma 4.1 to deduce that this solution is unique.) By Remark 4.1, the estimates of §§ 2 and 3 are valid for this solution. Hence a subsequence is convergent to a global solution of (2.1)-(2.10).

5. Asymptotic behavior.

THEOREM 5.1. *As* $t \to \infty$

$$u(x, t) \to u_a, \tag{5.1}$$

$$w(x, t) \to u_a \tag{5.2}$$

uniformly with respect to $x \in \Omega$.

Proof. From (3.1) and Poincaré's inequality we deduce that

$$\iint_{\Omega_\infty} \tilde{u}^2 < \infty \tag{5.3}$$

and from (3.10) we deduce that

$$|\tilde{u}(x, t) - \tilde{u}(x^1, t)| \leqq C|x - x^1|^{1/2} \tag{5.4}$$

for all x, x^1 in Ω, $t > 0$.

Suppose (5.1) is not true. Then there exist a sequence (x_n, t_n) in Ω_∞ with $t_n \to \infty$ and $c_0 > 0$, such that

$$|\tilde{u}(x_n, t_n)| \geqq c_0 > 0 \qquad (\tilde{u} = u - u_a).$$

By (5.4) it then follows that

$$\int_\Omega \tilde{u}^2(x, t_n)\, dx \geqq c > 0. \tag{5.5}$$

Let $t \in (t_n, t_n + 1)$. Then

$$|\tilde{u}(x, t) - \tilde{u}(x, t_n)| \leqq \int_{t_n}^{t} |u_\tau(x, \tau)|\, d\tau \leqq \left[\int_{t_n}^{t} u_t^2(x, \tau)\, d\tau\right]^{1/2}.$$

Hence

$$\int_\Omega |\tilde{u}(x, t) - \tilde{u}(x, t_n)|^2\, dx \leqq \int_{t_n}^{t_n+1} \int_\Omega u_t^2(x, t)\, dt \equiv \varepsilon_n \to 0 \tag{5.6}$$

if $n \to \infty$, by (3.10).

By the triangle inequality

$$\int_\Omega \tilde{u}^2(x, t_n)\, dx \leqq 2\int_\Omega \tilde{u}^2(x, t)\, dx + 2\int_\Omega |\tilde{u}(x, t) - \tilde{u}(x, t_n)|^2\, dx.$$

Recalling (5.5), (5.6) we then get

$$\int_\Omega \tilde{u}^2(x, t) \geqq \frac{c}{2} - 2\varepsilon_n \geqq \frac{c}{4} \qquad (t_n \leqq t \leqq t_n + 1),$$

if n is sufficiently large. Hence

$$\int_{t_n}^{t_n+1}\int_\Omega \tilde{u}^2(x, t)\, dx\, dt \geqq \frac{c}{4},$$

which is a contradiction to (5.3).

Having proved (5.1), the assertion (5.2) follows by solving

$$\tilde{w}_x + \tilde{w} = \tilde{u} \quad \text{with } \tilde{w}(-N, t) = 0$$

and using (5.1).

COROLLARY 5.2. *If $u_a < u_c$, then there exists a $T_0 > 0$ such that $r = 0$ if $t > T_0$.*

Indeed, by Theorem 5.1, $|u - u_a| < u_c - u_a$ if t is large enough, say $t > T_0$; this implies that $H(u - u_c) = 0$ and thus also $r = 0$ if $t > T_0$.

Corollary 5.2 means that the combustion has died out after a finite time T_0.

6. The Cauchy problem. In this section we extend the results of §§ 4 and 5 to the Cauchy problem, i.e., to the case $N = \infty$. We assume that

(6.1) $\sigma_0(x)$, $u_0(x)$ are positive-valued Holder continuous functions in $\mathbb{R}^1$,

(6.2) $$0 < \sigma_* \leqq \sigma_0(x) < 1 \quad \text{if } |x| > R_0 \text{ for some } R_0 > 0,$$

(6.3) $$u_0(x) \to u_a \quad \text{if } |x| \to \infty,$$

(6.4) $$\int_{\mathbb{R}} \sigma_0(x)(u_0(x) - u_a)^2\, dx < \infty,$$

and

(6.5) $$\int_{\mathbb{R}} k^2(u_0)(u_{0,x}(x))^2\, dx < \infty.$$

THEOREM 6.1. *Assume that* (2.12), (2.13), *and* (6.1)-(6.5) *hold. Then there exists a solution (σ, w, u, g) of* (2.1)-(2.7) *in $\mathbb{R}^1 \times (0, \infty)$ satisfying*

(6.6) $$w(x, t) \to u_a \quad \text{if } x \to -\infty, \quad t > 0,$$

(6.7) $$u(x, t) \to u_a \quad \text{if } |x| \to \infty, \quad t > 0,$$

(6.8) $$g(x, t) = g_a \quad \text{if } x > -R_0, \quad t > 0.$$

Proof. We modify the definition of $u_0(x)$ for $|x| > N - 1$ so that the new function $u_{0,N}(x)$ satisfies

(6.9) $$u_{0,N}(\pm N) = u_a, \qquad u_{0,N}(x) \geqq u_* > 0,$$

it is Holder continuous, and

(6.10) $$\int_{|x|<N} \sigma_0(x)(u_{0,N}(x) - u_a)^2\, dx \leqq C,$$

(6.11) $$\int_{|x|<N} k^2(u_{0,N})\left(\frac{\partial}{\partial x} u_{0,N}(x)\right)^2 dx \leqq C$$

where C is a positive constant independent of N.

Denote by $(\sigma_N, w_N, u_N, g_N)$ the solution (2.1)-(2.10) established in Theorem 4.2 for $\Omega = \Omega_N \equiv \{x; -N < x < N\}$ and $u_0 = u_{0,N}$. From (6.2) we have

$$r_N(x, t) = 0 \quad \text{if } |x| > R_0 \tag{6.12}$$

where r_N is the reaction rate r corresponding to the parameter N. Next, by comparison (since $w_N + r_N \geqq 0$), we obtain, for $\alpha > 1/(\sup \sigma_0(x))$,

$$u_N(x, t) \geqq u_* e^{-\alpha t}. \tag{6.13}$$

Also, from (2.2) by integration,

$$w_N(x, t) \geqq u_a e^{-x-N} + u_* e^{-\alpha t} \int_{-N}^{x} e^{-x+y}\, dy$$

so that

$$w_N(x, t) \geqq u_* e^{-\alpha t}. \tag{6.14}$$

Observe now that the proof of Lemma 3.1 is valid with C independent of N, since

$$\int_{\Omega_N} \int_0^t r_N\, dt\, dx \leqq C_1$$

where C_1 is independent of N (by (6.12)), and since (6.10) holds.

Next,

$$\begin{aligned} |\tilde{w}_N(x, t)| &= \int_{-N}^{x} e^{-x+y} |\tilde{u}_N(y, t)|\, dy \leqq \left[\int_{-N}^{x} e^{-2x+2y}\, dy\right]^{1/2} \left[\int_{-N}^{x} \tilde{u}_N^2(y, t)\, dy\right]^{1/2} \\ &\leqq \left[\int_{-N}^{x} \tilde{u}_N^2(y, t)\right]^{1/2} \leqq C \quad \text{by (3.1)} \end{aligned} \tag{6.15}$$

with C independent of N. This yields (3.5) and thus also (3.6), with C independent of N. The proof of (3.7) (with C independent of N) then follows as before.

Using (6.11), the proof of Lemma 3.3 follows as before with C independent of N.

Having established all the estimates of Lemmas 3.1-3.3 with C independent of N, we also obtain the estimates of Lemmas 3.4 and 3.5; thus

$$\int_0^T \int_D \left[\left|\frac{\partial}{\partial x} u_N\right|^p + \left|\frac{\partial^2}{\partial x^2} u_N\right|^p\right] dx\, dt \leqq C$$

for any bounded domain D in $\mathbb{R}^1$ where C is a constant depending on D and T but independent of N.

We can now take a subsequence of solutions $(\sigma_N, w_N, u_N, g_N)$ which converge uniformly on compact sets to a solution (σ, w, u, g) of (2.1)-(2.7). The assertion (6.8) follows from (6.12) and (2.4), (2.10). From (6.15) we get

$$|\tilde{w}(x, t)| \leqq \int_{-\infty}^{x} (\tilde{u}(y, t))^2\, dy \to 0 \quad \text{if } x \to -\infty,$$

and from

$$\int [\tilde{u}^2(x, t) + (\tilde{u}_x(x, t))^2]\, dx < \infty \quad \forall t > 0$$

we deduce (as in the proof of (5.1)) that

$$\tilde{u}(x, t) \to 0 \quad \text{if } |x| \to \infty \quad \forall t > 0.$$

Thus the proof of Theorem 6.1 is complete.

Remark 6.1. The solution constructed in Theorem 6.1 satisfies the integral estimates of Lemmas 3.1 and 3.3 (with $\Omega = \mathbb{R}^1$) and u_x, u_t, u_{xx} belong to $L^p_{\text{loc}}(\mathbb{R}^1 \times (0, \infty))$ for any $1 < p < \infty$.

Remark 6.2. We do not expect Theorem 5.1 to extend to the present case (of $\Omega = \mathbb{R}^1$) since, in general, a solution of

$$w_t - w_{xx} = f(x, t) \qquad \left(\int_0^\infty \int_{\mathbb{R}} f^2(x, t)\, dx\, dt < \infty \right),$$

$$w(x, 0) \to 0 \quad \text{if } |x| \to \infty, \qquad \int_{\mathbb{R}} w^2(x, 0)\, dx < \infty,$$

does not converge to zero as $t \to \infty$, x bounded. Indeed, a simple example is given by

$$w(x, t) = \exp\left\{ -\frac{x^2}{4(t+1)} \right\}.$$

REFERENCES

[1] R. R. BAKER, *Product formation mechanisms inside a burning cigarette*, Progr. Energy and Combust. Sci., 7 (1981), pp. 135-153.

[2] D. A. FRANK-KAMENETSKII, *Diffusion and Heat Transfer in Chemical Kinetics*, Plenum Press, New York, 1969.

[3] D. A. LAWSON AND J. NORBURY, *Numerical Methods in Thermal Problems*, Vol. III, R. W. Lewis, ed., Pineridge, Swansea, 1983.

[4] ———, *Numerical Methods in Heat Transfer*, Vol. III, R. W. Lewis, ed., Wiley, Chichester, 1985.

[5] J. NORBURY AND A. M. STUART, *Models for porous medium combustion*, to appear.

[6] T. J. OHLEMILLER, J. BELLAN, AND F. ROGERS, *A model of smouldering combustion applied to flexible polyurethane foams*, Combustion and Flame, 35 (1979), pp. 197-215.

[7] M. W. THRING, *The Science of Flames and Furnaces*, Chapman and Hall, London, 1962.

[8] W. R. WADE, J. E. WHITE, AND J. J. FLOREK, *Diesel particular trap regeneration techniques*, Technical Paper 810118, Society of Automative Engineers, 1981.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

002

GLOBAL EXISTENCE FOR A MODEL OF ELECTROPHORETIC SEPARATION*

JOEL D. AVRIN†

Abstract. We examine the modeling equations for a particular electrophoretic separation technique known as isotachophoresis. These equations form a system of advection-diffusion type and describe the time evolution of a number of charged chemical species. The transport mechanism arises from an electric field E where E_x is a superposition of the species concentrations; thus the equations are nonlinear. The spatial domain is the real line and the concentrations satisfy Dirichlet boundary conditions at $\pm\infty$. We show that these equations have global strong solutions that are unique in an appropriate sense.

Key words. global existence, isotachophoresis, advection-diffusion equation

AMS(MOS) subject classification. 35K

1. Introduction. The theory of electrophoretic separation describes the movement of charged particles in solution exposed to an electric field. Applications include the development of techniques for separating protein and other biological materials [2].

The strength of the electric field depends on the levels of concentrations of the species of charged particles present in the solution. Thus the modeling equations for the concentration levels and the field are nonlinear. Noting that the chemical reactions of the species occur on much shorter time scales than that of the diffusion or field-induced transport mechanisms, it is usually assumed that these reactions are in equilibrium; moreover, the experimental setup often admits the assumption that spatial variation is essentially one-dimensional. Under these hypotheses, the equations of electrophoresis, as developed in [6], are

$$(u_i)_t = [z_i\Omega_i E u_i + d_i(u_i)_x]_x, \qquad i = 1, \cdots, m, \tag{1.1a}$$

$$\varepsilon E_x = -e \sum_{k=1}^{m} z_k u_k. \tag{1.1b}$$

Here, for $u = u(x, t)$, u_x denotes $\partial u/\partial x$ and u_t denotes $\partial u/\partial t$. The unknowns in (1.1) are the species concentrations $u_i = u_i(x, t)$ and the electric field $E = E(x, t)$. The other parameters are known constants: Ω_i are the ionic mobilities and d_i are the diffusivities corresponding to each species, while $z_i = +1$ or -1 depending, respectively, on whether the u_i represent concentrations of positive or negative ions. In (1.1b) e is the molar charge and ε is the permittivity of the solvent.

In this paper we will establish the existence and uniqueness of global strong solutions of (1.1) with boundary conditions appropriate for a particular extensively used separation technique known as isotachophoresis, or ITP. In the usual set-up for ITP, the reaction column is long and is connected at both ends to large electrolyte reservoirs that negate the influence of reactions occurring at the electrodes. This makes the concentrations constant at the column ends, and effectively renders the system infinitely long [3], [6]. Thus x varies over the entire real line in (1.1), and the concentrations satisfy fixed Dirichlet boundary conditions:

$$u_i(-\infty) = \alpha_i, \qquad u_i(+\infty) = \beta_i. \tag{1.1c}$$

* Received by the editors January 21, 1987; accepted for publication May 20, 1987.
† Department of Mathematics, University of North Carolina, Charlotte, North Carolina 28223.

The boundary conditions for the electric field E can be deduced from (1.1c) and the fact that the electric current I through the medium is constant.

Setting

$$f_i = z_i\Omega_i E u_i + d_i(u_i)_x, \tag{1.2}$$

we have, as noted in [3], that

$$I = e\sum_{i=1}^{m} z_i(f_i)(-\infty) = e\sum_{i=1}^{m} z_i(f_i)(+\infty). \tag{1.3}$$

Since u_i must also satisfy

$$(u_i)_x(-\infty, t) = (u_i)_x(+\infty, t) = 0 \tag{1.4}$$

(see e.g. [3]) we have, from (1.1c) and (1.4) that (1.3) reduces to

$$I = e\sum_{i=1}^{m} \Omega_i\alpha_i E(-\infty, t) = e\sum_{i=1}^{m} \Omega_i\beta_i E(+\infty, t). \tag{1.5}$$

Since I is a constant, we must have from (1.5) that $E(\pm\infty, t) = E_{\pm}$ where E_+ and E_- are constants. We will see in § 2 that E can be determined in a completely satisfactory way by (1.1a)-(1.1c), (1.5), and the initial conditions

$$u_i(x, 0) = u_i^0. \tag{1.6}$$

Finally, as in [3], we assume that

$$\sum_{i=1}^{m} z_i\alpha_i = \sum_{i=1}^{m} z_i\beta_i = 0. \tag{1.7}$$

Condition (1.7) is a natural condition to impose in the applications in light of the separation mechanism and it will play an important role in what follows.

In § 2 below, we make some preliminary observations which will allow us to recast problem (1.1a)-(1.1c) together with the conditions (1.4)-(1.7) as a system of m equations of advection-diffusion type. Section 2 will conclude with the statement of our main local existence theorem. In § 3 we will prove this theorem, and in § 4 we will show that our solutions are global in time.

We note that the global existence result of this paper can be viewed as a companion result to the development in [3], in which traveling wave solutions were found for (1.1a)-(1.1c) with the conditions (1.4), (1.5), (1.7).

2. Some preliminary observations. We first define auxiliary functions $w_i = w_i(x)$ as follows:

$$w_i(x) = \begin{cases} \alpha_i, & x \leqq -1, \\ e_i(x), & -1 \leqq x \leqq 1, \\ \beta_i, & x \geqq 1, \end{cases} \tag{2.1}$$

where e_i are smooth functions that join the horizontal portions of w_i in a C^∞ manner, such that $\alpha_i \leqq e_i(x) \leqq \beta_i$ and such that

$$(w_i)_x \in C_0^\infty \tag{2.2}$$

where C_0^∞ denotes $C_0^\infty(\mathbb{R})$. Note from (1.7) that we also have

$$\sum_{i=1}^{m} z_i w_i \in C_0^\infty. \tag{2.3}$$

We can assume that the solutions u_i of (1.1) have the form

$$u_i = v_i + w_i. \tag{2.4}$$

where for each t $v_i(\cdot, t) \in W^{2,1} = W^{2,1}(\mathbb{R})$; the suitability of (2.4) proceeds from the smooth asymptotic behavior of u_i demonstrated in [3]. Note also that $v_i \in W^{2,1}$ implies that u_i and $(u_i)_x$ are in C_0, so that the right-hand side of (2.4) satisfies (1.1c) and (1.4). Eventually we will plug the right-hand side of (2.4) into (1.1) to obtain a system of equations in v_i. We first determine the form of E using (1.4)-(1.7). Assuming (2.4) we note that

$$U(x, t) \equiv \sum_{k=1}^{m} z_k u_k \in L^1 = L^1(\mathbb{R}) \tag{2.5}$$

where we have used (2.3). It is thus appropriate to set

$$E = -(e/\varepsilon) \int_{-\infty}^{x} U(y, t)\, dy + C(t) \tag{2.6}$$

and then solve for $C(t)$ using the boundary conditions. From (1.5) it is easy to see that

$$C(t) = \left[\sum_{k=1}^{m} \Omega_k \beta_k (-e/\varepsilon) \int_{-\infty}^{\infty} U(y, t)\, dy \right] \Big/ \left[\sum_{k=1}^{m} \Omega_k (\alpha_k - \beta_k) \right], \tag{2.7}$$

which is obtained simply by setting the two expressions on the right-hand side of (1.5) equal to each other. We now want to show that $C(t)$ is a *constant* which we will denote by E_-. Note from (1.1) that from superposition U satisfies

$$U_t = \left[\sum_{k=1}^{m} z_k f_k \right]_x \tag{2.8}$$

and thus

$$\frac{d}{dt} \int_{-\infty}^{\infty} U(y, t)\, dy = \sum_{k=1}^{m} z_k f_k(+\infty) - \sum_{k=1}^{m} z_k f_k(-\infty) = 0. \tag{2.9}$$

Hence

$$E(x, t) = -(e/\varepsilon) \int_{-\infty}^{x} U(y, t)\, dy + E_- \tag{2.10}$$

where E_- is the right-hand side of (2.7) with $U(y, t)$ replaced by

$$U(y, 0) = \sum_{k=1}^{m} z_k (u_k^0). \tag{2.11}$$

With E thus determined in terms of u_i and u_i^0, we are now ready to formulate our problem in terms of the u_i; setting $c_i = z_i \Omega_i$ and plugging $v_i + w_i$ into (1.1a) and (1.1b) in place of u_i, we obtain the following system of initial-value problems:

$$(v_i)_t = d_i (v_i)_{xx} + (c_i E)(v_i)_x + (c_i E_x) v_i + (c_i E)(w_i)_x + (c_i E_x) w_i + d_i (w_i)_{xx}, \tag{2.12a}$$

where E is given by (2.10) with

$$v_i(x, 0) = v_i^0 = u_i^0 - w_i(x), \tag{2.12b}$$

$$U(y, t) = \sum_{k=1}^{m} z_k (v_k + w_k). \tag{2.13}$$

Note that by (2.10) the system (2.12) is an integrodifferential equation. We will see in the next section that (2.12) can be treated using methods involving semigroups and variation-of-parameters formulas. Thus in the next section we will prove the following.

THEOREM 2.1. *The system* (2.12) *has a unique local strong solution* $v=(v_1, \cdots, v_m)$ *where each* $v_i \in C^1([0, T]; W^{2,1})$ *for some* $T>0$.

3. Proof of Theorem 2.1. We will let $\|\cdot\|_1$, $\|\cdot\|_{m,1}$ and $\|\cdot\|_\infty$ denote the norms on L^1, $W^{m,1}$, and C_0, respectively, where m is a positive integer. Let $M>0$ and $T>0$, with T to be determined. Let A_i be defined by $A_i f = d_i f_{xx}$ and let $W_i(t)$ denote the analytic semigroup generated by A_i. Let

$$Q=\Bigg\{u=(u_1, u_2, \cdots, u_m): u_i \in C([0, T]; W^{2,1}) \text{ and } \sup_i \sup_{0\le t\le T} \Bigg\| u_i(t) - W_i(t)u_i^0 - \int_0^t W_i(t-s)(w_i)_{xx}\, ds \Bigg\|_{2,1} \le M\Bigg\}. \tag{3.1}$$

If for each $u, v \in Q$ we set

$$\rho(u, v) = \sup_i \sup_{0\le t\le T} \| u_i(t) - v_i(t) \|_{2,1} \tag{3.2}$$

then (Q, ρ) is a complete metric space. For each i set

$$\begin{aligned}(S_i v_i)(t) &= W_i(t)v_i^0 + \int_0^t W_i(t-s)(w_i)_{xx}\, ds \\ &\quad + \int_0^t W_i(t-s)[c_i E(v(s))(v_i(s)+w_i)]_x\, ds\end{aligned} \tag{3.3}$$

where $E(v(s))$ is defined by (2.10) with $U(y, s)=\sum z_k(v_k+w_k)$. If we now let $Sv = (S_1v_1, S_2v_2, \cdots, S_mv_m)$, then we want to select T so that S is a contraction on E.

We first note that there exists a constant c such that

$$\| (W(t)f)_x \|_1 \le ct^{-1/2}\|f\|_1 \tag{3.4}$$

for all $f \in L^1$ and $t>0$, and a constant K_1 such that

$$\|f\|_\infty \le K_1 \|f\|_{1,1} \tag{3.5}$$

for all $f \in W^{1,1}$. Meanwhile, if $v \in E$ then

$$\begin{aligned}\| c_i E(v(s)) \|_\infty &\le |E_-| + \sup_x |c_i|\,|e/\varepsilon| \sum_{k=1}^m \int_{-\infty}^x (|v_k(y, s)| + |w_k(y)|)\, dy \\ &\le |E_-| + m\|\textstyle\sum w_k\|_1 \sup_i |c_i|\,|e/\varepsilon| \sup_k \|v_k(s)\|_1 \\ &\le |E_-| + m\|\textstyle\sum w_k\|_1 \sup_i |c_i|\,|e/\varepsilon| M \equiv K_2\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}\| c_i[E(v(s))]_x \|_1 &\le |c_i|\,|e/\varepsilon| \Bigg(\sum_{k=1}^m \|v_k(s)\|_1 + \|\textstyle\sum w_k\|_1 \Bigg) \\ &\le \sup_i |c_i|\,|e/\varepsilon|\, \|\textstyle\sum w_k\|_1 mM \equiv K_3.\end{aligned} \tag{3.7}$$

If we set

$$G_i(t) = W_i(t)v_i^0 + \int_0^t W_i(t-s)(w_i)_{xx}\,ds \tag{3.8}$$

then from (3.3)-(3.8) we have

$$\begin{aligned}
&\|[(S_iv_i)(t) - G_i(t)]_x\|_1 \\
&\quad \leqq \int_0^t c(t-s)^{-1/2}\|[c_iE(v(s))(v_i(s)+w_i)]_x\|_1\,ds \\
&\quad \leqq \int_0^t c(t-s)^{-1/2}[\|c_iE(v(s))\|_\infty\|(v_i)_x(s)\|_1 \\
&\qquad + \|c_i[E(v(s))]_x\|_1\|v_i(s)\|_\infty + \|c_iE(v(s))\|_\infty\|(w_i)_x\|_1 \\
&\qquad + \|c_i[E(v(s))]_x\|_1\|w_i\|_\infty]\,ds \\
&\quad \leqq cT^{1/2}[(K_2+K_1K_3)M + K_2\|(w_i)_x\|_1 + K_3\|(w_i)\|_\infty].
\end{aligned} \tag{3.9}$$

In a similar fashion we can estimate

$$\|[(S_iv_i)(t) - G_i(t)]\|_1 \tag{3.10}$$

and

$$\|[(S_iv_i)(t) - G_i(t)]_{xx}\|_1. \tag{3.11}$$

We note that (3.10) will be bounded by the last line of (3.9) except that $cT^{1/2}$ is replaced by T. Meanwhile (3.11) is bounded by

$$\int_0^t c(t,s)^{-1/2}\|[c_iE(v(s))(v_i(s)+w_i)]_{xx}\|_1\,ds \tag{3.12}$$

which we can estimate by computing:

$$\begin{aligned}
&[c_iE(v(s))(v_i(s)+w_i)]_{xx} \\
&\quad = 2c_i[E(v(s))]_x(v_i(s)+w_i)_x + c_iE(v(s))(v_i(s)+w_i)_{xx} \\
&\qquad + c_i[E(v(s))]_{xx}(v_i(s)+w_i).
\end{aligned} \tag{3.13}$$

We now can proceed as in (3.9) once we observe that

$$\begin{aligned}
\|c_i[E(v(s))]_{xx}\|_1 &\leqq \sup_i|c_i|\,|e/\varepsilon|\sum_{k=1}[\|(v_i)_x(s)\|_1 + \|(w_k)_x\|_1] \\
&\leqq \sup_i|c_i|\,|e/\varepsilon|[mM + \textstyle\sum\|(w_k)_x\|_1].
\end{aligned} \tag{3.14}$$

From these remarks and the development in (3.19) we see that T can be selected so that S maps Q to Q.

To show that we can select T smaller, if necessary, so that for $u, v \in Q$

$$\rho(Su, Sv) \leqq K(T)\rho(u, v) \tag{3.15}$$

with $0 < K(T) < 1$ now requires more tedious calculations but no deeper than those made above. The main points of the development come from replacing $E(v(s))$ by $E(u(s)) - E(v(s))$ and $v_i(s)$ by $v_i(s) - u_i(s)$ in (3.6), (3.7), and (3.14).

These observations demonstrate that we have solutions $v_i(t)$ of the appropriate integral equations. To see that each v_i is a strong solution now proceeds in a standard fashion. We already have that $v_i(t)$ is in the domain of A_i for each t; we now differentiate under the integral sign in the usual manner (see,e.g.,[5]) using the dominated convergence theorem. This completes the proof of Theorem 2.1.

4. Global existence. Let $[0, T)$ be the maximal interval of existence for the solutions found in § 3. We first show that the L^1-norm of the v_i stays bounded on $[0, T)$, if T is finite. Consider, for $0 \leqq t < T$ and for each i, the *linear* equation

$$z_t = d_i z_{xx} + [c_i E z]_x, \tag{4.1a}$$

$$z(x, 0) = z_0, \tag{4.1b}$$

where z_0 is in $W^{2,1}$ and satisfies $z_0(x) \geqq 0$ for all x, while $E = E(v(t))$ with v as in § 3. It is clear that (4.1) has a solution $z(x, t)$ throughout $[0, T)$ with $z \in W^{2,1}$ for each t; in fact, the methods of § 3 apply; also note that $E \in C([0, T); C_B(\mathbb{R}))$. Therefore there exists a fundamental solution $U(t, \tau)$ such that $z(t) = U(t, 0)z_0$. It is clear, by applying the Trotter product formula for fixed τ, that $U(t, \tau)$ is positivity preserving. Moreover, since $z(t) \in W^{2,1}$,

$$\frac{d}{dt}\int_{-\infty}^{\infty} z(x, t)\, dx = \int_{-\infty}^{\infty} [c_i E z]_x\, ds = 0; \tag{4.2}$$

hence $U(t, \tau)$ preserves the L^1-norm for nonnegative initial data. For arbitrary initial data $z_0 \in W^{2,1}$, set $z_0^+ = \max\{z_0, 0\}$ and $z_0^- = \max\{-z_0, 0\}$, then

$$\begin{aligned} \| U(t, \tau) z_0 \|_1 &\leqq \| U(t, \tau) z_0^+ \|_1 + \| U(t, \tau) z_0^- \|_1 \\ &= \| z_0^+ \|_1 + \| z_0^- \|_1 = \| z_0 \|_1; \end{aligned} \tag{4.3}$$

thus $U(t, \tau)$ is a *contraction* on L^1 for each $0 \leqq \tau \leqq t < T$.

By variation-of-parameters our solutions v_i evidently satisfy

$$\begin{aligned} v_i(t) = U(t, 0)v_i^0 &+ \int_0^t U(t, s)(c_i E)(w_i)_x\, ds \\ &+ \int_0^t U(t, s)(c_i E_x)(w_i)\, ds \\ &+ \int_0^t U(t, s)\, d_i (w_i)_{xx}\, ds; \end{aligned} \tag{4.4}$$

hence for $0 \leqq t \leqq T$

$$\begin{aligned} \| v_i(t) \|_1 &\leqq \| v_i^0 \|_1 + \int_0^t \| c_i E \|_\infty \| (w_i)_x \|_1\, ds \\ &\quad + \int_0^t \| c_i E_x \|_1 \| w_i \|_\infty\, ds + d_i T \| (w_i)_{xx} \|_1 \\ &\leqq \| v_i \|_1 + d_i T \| (w_i)_{xx} \|_1 \\ &\quad + [|E_-| + \sup_i |c_i|\,|e/\varepsilon|][\| (w_i)_x \|_1 + \| w_i \|_\infty] T \| \textstyle\sum w_k \|_1 \\ &\quad + [|E_-| + \sup_i |c_i|\,|e/\varepsilon|][\| (w_i)_x \|_1 + \| w_i \|_\infty] \\ &\quad \cdot \int_0^t \sum_{k=1}^m \| v_k(s) \|_1\, ds \end{aligned} \tag{4.5}$$

where we have used (3.6) and (3.7). If we now sum up (4.5) for each i, we see that by Gronwall's inequality

$$\sum_{i=1} \| v_i(t) \|_1 \leqq K_6 \exp(K_7 T) \equiv N \tag{4.6}$$

for all $t \in [0, t)$, where K_6 and K_7 are constant on $[0, T)$.

We now use (4.6) to estimate $\| v_i \|_{1,1}$ on $[0, T)$. From the integral equation representation we have, for each $0 \leqq t < T$,

$$\begin{aligned}
\| v_i(t) \|_{1,1} &\leqq N + \| (v_i)_x(t) \|_1 \\
&\leqq N + \| (v_i^0)_x \|_1 + \int_0^t c(t-s)^{-1/2} \| (w_i)_{xx} \|_1 \, ds \\
&\quad + \int_0^t c(t-s)^{-1/2} [\| c_i E \|_\infty \| (v_i)_x \|_1 + \| c_i E_x \|_1 \| v_i \|_\infty \\
&\qquad + \| c_i E \|_\infty \| (w_i)_x \|_1 + \| c_i E_x \|_1 \| w_i \|_\infty] \, ds \qquad (4.7) \\
&\leqq N + cT^{1/2} \| (w_i)_{xx} \|_1 + \| (v_i^0)_x \|_1 + K_8 T (\| (w_i)_x \|_1 + \| w_i \|_\infty) \\
&\quad + K_8 \int_0^t c(t-s)^{-1/2} (1 + K_1) \| v_i(s) \|_{1,1} \, ds
\end{aligned}$$

where

$$K_8 = |E_-| + \sup_i |c_i| \, |e/\varepsilon| (N + \| \textstyle\sum w_k \|_1) \tag{4.8}$$

so again by Gronwall's inequality $\| v_i(t) \|_{1,1}$ stays bounded on $[0, T)$. Note that we have used (3.6), (3.7), and (4.6) to obtain (4.7). We can now use the uniform bound on $\| v_i(t) \|_{1,1}$ to show that $\| v_i(t) \|_{2,1}$ stays bounded on $[0, T)$; the development is similar to that used in (4.7), but now we also use (3.14).

Thus, if $T < +\infty$, standard arguments now obtain a contradiction in the usual fashion (see, e.g., [4], [5]). Hence we have demonstrated the following.

THEOREM 4.1. *The solutions found in* § 3 *are* <u>*global*</u> *solutions, i.e.,* $v_i \in C([0, +\infty); W^{2,1}) \cap C^1([0, +\infty); L^1)$ *for each* i.

5. Remarks. We have shown that (2.12) has global strong solutions in $W^{2,1}$ for each t. Using standard bootstrap techniques, however, we can show that our solutions lie in $\cap_n W^{n,1}$ for each $t > 0$; in particular each $v_i(t)$ is C^∞ for positive t. Our techniques of the last section to pass from bounds on the $W^{1,1}$ norm to bounds on the $W^{2,1}$ norm using (3.14) can be applied inductively to obtain this expected parabolic regularity.

Finally, there is an additional a priori estimate that can be derived for the solutions v_i. Rewriting (2.12a) as

$$(v_i)_t = [c_i E (v_i + w_i) + d_i (v_i)_x]_x \tag{5.1}$$

we integrate both sides of (5.1) to see that

$$\frac{d}{dt} \int_{-\infty}^{\infty} v_i(y, t) \, dy = [c_i E w_i]_{-\infty}^{\infty}, \tag{5.2}$$

so there exists a constant B_i, depending only on α_i, β_i, and $E_\pm$ such that

$$\int_{-\infty}^{\infty} v_i(y, t) \, dy = B_i t + \int_{-\infty}^{\infty} v_i^0(y) \, dy. \tag{5.3}$$

Since v_i is not necessarily nonnegative, it is not clear what the usefulness of (5.3) is, but perhaps it may prove to have some physical significance.

Acknowledgment. I would like to thank Professor Paul Fife of the University of Arizona for suggesting this problem to me and for many useful discussions.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] Z. DEYL, ed., *Electrophoresis: A Survey of Techniques and Applications*, Elsevier, Amsterdam, 1979.
[3] P. C. FIFE, O. A. PALUSINSKI, AND V. SU, *Electrophoretic traveling waves*, Trans. Amer. Math. Soc., to appear.
[4] J. A. GOLDSTEIN, *Semigroups of Linear Operators and Applications*, Oxford University Press, New York, Oxford, 1985.
[5] M. REED, *Abstract Nonlinear Wave Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1975.
[6] D. A. SAVILLE AND O. A. PALUSINSKI, *Theory of electrophoretic separation, Parts* 1, 2, AIChE J., 32 (1986).

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

003

REGULARITY AND STRONG CONVERGENCE OF A VARIATIONAL APPROXIMATION TO A NONHOMOGENEOUS DIRICHLET HYPERBOLIC BOUNDARY VALUE PROBLEM*

IRENA LASIECKA† AND JAN SOKOLOWSKI‡

Abstract. Variational approximations of the second order scalar hyperbolic equations with nonhomogeneous Dirichlet boundary data are considered. Strong convergence of the approximate solutions to the original ones is established.

Key words. variational approximations, hyperbolic equations, Dirichlet boundary conditions, regularity of solutions

AMS(MOS) subject classification. 35L20

1. Introduction. Let Ω be an open bounded domain in R^n with a smooth boundary Γ. Consider the following second order scalar hyperbolic equation:

$$\begin{aligned} &\ddot{u}(t,x)=\Delta u(t,x)+f(t,x), \qquad t,x\in(0,T]\times\Omega\equiv Q,\\ &u(0,x)=\dot{u}(0,x)=0 \quad \text{in } \Omega,\\ &u(t,x)|_{\Gamma}=g(t,x) \quad \text{in } (0,T)\times\Gamma\equiv\Sigma. \end{aligned} \tag{1.1}$$

Optimal regularity properties of the solution u, which improve upon previous literature, e.g., [LM], are established in [L1], [LT1], [LLT]; in particular, the following estimate holds:

$$|u|_{C[0T;L_2(\Omega)]}\leqq C[|g|_{L_2(\Sigma)}+|f|_{L_1[0,T;[H_0^1(\Omega)]']}].$$

The following question may be asked: how do we construct a numerical algorithm in order to compute effectively u from the boundary data g? It is well known that the "best" numerical approximations of various partial differential equation problems are based on a certain variational formulation of the original equation. The problem in our case is, however, that due to the Dirichlet nature of the nonhomogeneous boundary condition problem (1.1) does not admit a natural variational formulation (in contrast, a natural variational formulation is standard in the case of Neumann or Robin boundary conditions). In this context, the idea of Lions [L1] is to "approximate" the solution $u(t)$ of (1.1) by a sequence of functions $u_\varepsilon(t)$ which are determined as solutions of the following problems:

$$\begin{cases} \ddot{u}_\varepsilon(t,x)=\Delta u_\varepsilon(t,x)+f(t,x) & \text{in } Q,\\ u_\varepsilon(0,x)=\dot{u}_\varepsilon(0,x)=0 & \text{in } \Omega,\\ \varepsilon\dfrac{\partial u_\varepsilon}{\partial\eta}+\beta u_\varepsilon=\beta g & \text{in } \Sigma \end{cases} \tag{1.2}$$

where β is a self-adjoint second order elliptic operator defined on the variety Γ.[1] For example, one can take $\beta\equiv-\Delta_\Gamma+1$ where Δ_Γ is the Laplace's Beltrami operator corresponding to the Riemann metric on Γ induced by R^n.

* Received by the editors June 16, 1986; accepted for publication (in revised form) April 12, 1987.

† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22903. The research of this author was supported in part by the National Science Foundation under grant DMS-8301668 and by the Air Force Office of Scientific Research under grant AFOSR-84-0365.

‡ Systems Research Institute, Polish Academy of Sciences, Warsaw, ul Nevelske 6, Poland. The research of this author was completed while the author was visiting the Department of Mathematics, University of Florida, Gainesville, Florida 32611.

[1] This in particular implies that $\beta: H^{s+2}(\Gamma)\to H^s(\Gamma)$ is an isomorphism.

The advantage of introducing (1.2) is, of course, that (1.2) admits a natural variational formulation

$$(1.3)\qquad \begin{cases} (\ddot{u}_\varepsilon, \phi)_\Omega + (\nabla u_\varepsilon, \nabla \phi)_\Omega + \dfrac{1}{\varepsilon}\langle \beta u_\varepsilon, \phi\rangle_\Gamma = \dfrac{1}{\varepsilon}\langle \beta g, \phi\rangle_\Gamma + (f, \phi)_\Omega, & \phi \in C^2(\bar{\Omega}), \\ u_\varepsilon(0) = \dot{u}_\varepsilon(0) = 0. \end{cases}$$

In [L1] it was shown that $u_\varepsilon(t)$ approximates $u(t)$ in the following sense: for any $g \in L_2(\Sigma)$ and $f \in L_1[0, T; L_2(\Omega)]$

$$(1.4)\qquad u_\varepsilon \to u \text{ in } L^\infty[0, T; L_2(\Omega)] \text{ weak star, when } \varepsilon \to 0.$$

In view of the above, we can think of (1.1) as a limit problem for (1.2). Therefore, in order to find an effective numerical approximation of (1.1), the natural idea to pursue is to look for numerical algorithms (Ritz-Galerkin, finite element, etc.) of the variational equality (1.3). However, in order to establish the convergence or even more—the rates of the convergence—of these approximations, a necessary prerequisite is to know more about the regularity properties of the solutions to (1.2) as well as their convergence to $u(t)$. Thus, the main purpose of the present paper is to study regularity (more precisely uniform differentiability) properties of the solutions $u_\varepsilon(t)$ along with the convergence of $u_\varepsilon(t)$ to $u(t)$. In particular, we will prove that the convergence in (1.4) is, in fact, strong. We shall also establish a number of regularity results for $u_\varepsilon(t)$, which are reminiscent of those valid for the limit solution $u(t)$. These results, besides being of interest in their own right, are of fundamental importance in the study of numerical schemes approximating (1.2). In fact, they are used crucially in [LS], where finite element techniques are developed to approximate $u_\varepsilon(t)$ and hence $u(t)$.

The outline of the paper is as follows. In § 2 we provide some background material on the properties of the solution $u(t)$ and we state our main results, Theorems 2.3 and 2.4. In § 3 we discuss the regularity and convergence of the steady state solutions to (1.2). These results are needed for § 5 where the proof of our main result, Theorem 2.4, on the uniform differentiability of the solution $u_\varepsilon(t)$ is provided. Section 4, instead, is devoted to the proof of Theorem 2.3 which states convergence of $u_\varepsilon(t)$ to $u(t)$.

Notation.

$(\,,\,)$ and $\|\ \|$ stand for inner product and the norm in $L_2(\Omega)$.

$\langle\,,\rangle$ and $|\ |$ stand for inner product and the norm in $L_2(\Gamma)$.

$H^r(\Omega)$, $H^{r,s}(Q)$ (respectively, $H^r(\Gamma)$, $H^{r,s}(\Sigma)$) denote the usual Sobolev space defined as in [LM].

$H^{-r,-s}(Q) \equiv (H^{r,s}(Q))'$.

$H^{-r}(\Omega) = (H^r(\Omega))'$; $H^{-r}(\Gamma) = H^r(\Gamma)'$.

$H_0^{-r}(\Omega) \equiv (H_0^r(\Omega))'$.

$\mathscr{L}(X \to Y)$ is the space of linear, bounded transformations from Banach space X to another Banach space Y.

$\dot{u} = (d/dt)u$; $\ddot{u} = d^2/dt^2$.

2. Preliminaries and the statements of the main results. Let us begin by collecting regularity results available for the original problem (1.1).

THEOREM 2.1 ([LT1], [LT2], [LLT], [L2], [S1]). *Let u be the solution to* (1.1) *with $g \in L_2(\Sigma)$ and $f = 0$. Then*

$$(2.1)\qquad |u|_{C[0T;L_2(\Omega)]} + |\dot{u}|_{C[0T;(H_0^1(\Omega))']} + \left|\frac{\partial u}{\partial \eta}\right|_{H^{-1,-1}(\Sigma)} \leqq C^2 |g|_{L_2(\Sigma)}.$$

[2] C will stand for a generic constant.

If in addition we assume that $g \in H^{1,1}(\Sigma)$, $g(0)=0$ *and we take* $f \in L_1[0T; L_2(\Omega)]$ *then*

$$|u|_{C[0T;H^1(\Omega)]}+|\dot{u}|_{C[0T;L_2(\Omega)]}+\left|\frac{\partial u}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C[|g|_{H^{1,1}(\Sigma)}+|f|_{L_1[0T;L_2(\Omega)]}]. \tag{2.2}$$

More generally with $g \in H^{s,s}(\Sigma)$, $s \geqq 1$ *where* g *satisfies the appropriate compatibility conditions and with* $f=0$ *we have*

$$|u|_{C[0T;H^s(\Omega)]}+|\dot{u}|_{C[0T;H^{s-1}(\Omega)]}+\left|\frac{\partial u}{\partial \eta}\right|_{H^{s-1,s-1}(\Sigma)} \leqq C|g|_{H^{s,s}(\Sigma)}. \tag{2.3}$$

Remark 2.1. Notice that the regularity of the solution on the boundary does not follow from the interior regularity. In fact, the regularity of the normal derivative of the solution on the boundary is higher than the Trace Theorem combined with interior regularity would imply.

Next, let u_ε stand for the solution to (1.2). The following results were proved in [L1].

THEOREM 2.2 [L1]. *Let* u (*respectively,* u_ε) *be the solution to* (1.1) (*respectively,* (1.2)) *with* $g=0$ *and* $f \in L_1[0T; L_2(\Omega)]$. *Then*

$$|u_\varepsilon|_{C[0T;H^1(\Omega)]}+|\dot{u}_\varepsilon|_{C[0T;L_2(\Omega)]}+\left|\frac{\partial y_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)}+|\dot{u}_\varepsilon|_{L_2(\Sigma)} \leqq C|f|_{L_1[0T;L_2(\Omega)]}, \tag{2.4}$$

$$|u_\varepsilon|_{C[0T;H^1(\Gamma)]} \leqq C\sqrt{\varepsilon}|f|_{L_1[0T;L_2(\Omega)]}. \tag{2.5}$$

(2.6) (i) $u_\varepsilon \to u$ *in* $L^\infty[0T; H^1(\Omega)]$ *weak star,*

(ii) $\dot{u}_\varepsilon \to \dot{u}$ *in* $L^\infty[0T; L_2(\Omega)]$ *weak star,*

(iii) $u_\varepsilon \to 0$ *in* $C[0T; H^1(\Gamma)]$,

(iv) $\dfrac{\partial u_\varepsilon}{\partial \eta} \to \dfrac{\partial u}{\partial \eta}$ *in* $L^2(\Sigma)$ *weakly.*

With $g \in L_2(\Sigma)$ *in* (1.1) (*respectively,* (1.2)) *and* $f \in L_1[0T; L_2(\Omega)]$ *we have*

$$|u_\varepsilon|_{C[0T;L_2(\Omega)]} \leqq C[|g|_{L_2(\Sigma)}+|f|_{L_1[0T;L_2(\Omega)]}], \tag{2.7}$$

$$u_\varepsilon \to u \quad \text{in } L^\infty[0T; L_2(\Omega)] \text{ weak star}, \tag{2.8}$$

where C *stand for a generic constant* <u>*independent*</u> *of* $\varepsilon > 0$.

The main goals of this paper are (i) to prove the uniform (with respect to ε) differentiability of the solutions u_ε with nonhomogeneous, smooth boundary data g, and (ii) to improve the convergence results of (2.6) and (2.8).

Our results are as follows.

THEOREM 2.3 (Differentiability). (i) *Let* u (*respectively,* u_ε) *be the solution to* (1.1) (*respectively,* (1.2)) *with* $g \in H^{1,1}(\Sigma)$; $g(0)=0$ *and* $f \equiv 0$. *Then*

$$|u_\varepsilon|_{C[0T;H^1(\Omega)]}+|\dot{u}_\varepsilon|_{C[0T;L_2(\Omega)]} \leqq C|g|_{H^{1,1}(\Sigma)}, \tag{2.9}$$

$$\left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)}+|\dot{u}_\varepsilon|_{L_2(\Sigma)}+|u_\varepsilon|_{L_2[0T;H^1(\Gamma)]} \leqq C|g|_{H^{1,1}(\Sigma)}, \tag{2.10}$$

$$|u_\varepsilon - g|_{L_2[0T;H^2(\Gamma)]} \leqq C\varepsilon|g|_{H^{1,1}(\Sigma)}. \tag{2.11}$$

(ii) *If in addition we assume that* $g \in H^{s,s}(\Sigma)$, $s \geqq 1$ *and* g *satisfies the appropriate compatibility conditions then*

$$|u_\varepsilon|_{C[0T;H^s(\Omega)]}+|\dot{u}_\varepsilon|_{C[0T;H^{s-1}(\Omega)]} \leqq C|g|_{H^{s,s}(\Sigma)}, \tag{2.12}$$

$$\left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{H^{s-1,s-1}(\Sigma)}+|\dot{u}_\varepsilon|_{H^{s-1,s-1}(\Sigma)}+|u_\varepsilon|_{L_2[0T;H^s(\Gamma)]} \leqq C|g|_{H^{s,s}(\Sigma)}, \tag{2.13}$$

$$|u_\varepsilon - g|_{L_2[0T;H^{s+1}(\Gamma)]} \leqq C\varepsilon|g|_{H^{s,s}(\Sigma)}. \tag{2.14}$$

THEOREM 2.4 (Convergence). (i) *Let u (respectively, u_ε) be the solution to* (1.1) (*respectively*, (1.2)) *with $f \in L_1[0T; L_2(\Omega)]$ and $g = 0$. Then*

$$(2.15)\quad \text{(i)}\quad |u_\varepsilon|_{L_2[0T;H^2(\Gamma)]} \leqq C\varepsilon |f|_{L_1[0T;L_2(\Omega)]},$$

$$\text{(ii)}\quad |u_\varepsilon - u|_{L_\infty[0T;L_2(\Omega)]} \leqq C\varepsilon |f|_{L_1[0T;L_2(\Omega)]},$$

$$(2.16)\quad \text{(i)}\quad u_\varepsilon \to u \quad \text{in } C[0T;H^1(\Omega)],$$

$$\text{(ii)}\quad \dot{u}_\varepsilon \to \dot{u} \quad \text{in } C[0T;L_2(\Omega)],$$

$$(2.17)\quad \text{(i)}\quad \dot{u}_\varepsilon|_\Gamma \to 0 \quad \text{in } L_2(\Sigma),$$

$$\text{(ii)}\quad \frac{\partial u_\varepsilon}{\partial \eta} \to \frac{\partial u}{\partial \eta} \quad \text{in } L_2(\Sigma).$$

(ii) *Let u (respectively, u_ε) be the solution to* (1.1) (*respectively*, (1.2)) *with $f = 0$. Then*

$$(2.18)\qquad |u_\varepsilon - u|_{C[0T;L_2(\Omega)]} \leqq C\varepsilon |g|_{H^{1,1}(\Sigma)},$$

$$(2.19)\qquad u_\varepsilon \to u \quad \text{in } C[0T; L_2(\Omega)] \quad \text{for any } g \in L_2(\Sigma).$$

Remark 2.2. Notice that the regularity results of u_ε stated in Theorem 2.3 are reminiscent of those in Theorem 2.1. In fact, (2.9), (2.10) reconstruct the regularity of the original solution $u(t)$ given by (2.1) and (2.2).

3. Convergence of the steady state solutions.

3.1. Statement of results. In order to prove our regularity and convergence results for the problem (1.2), we first need to establish similar results for the corresponding elliptic problem. More precisely, consider the following elliptic problems:

$$(3.1)\qquad \Delta v = f \quad \text{in } \Omega, \qquad v|_\Gamma = 0 \quad \text{in } \Gamma$$

and

$$(3.2)\qquad \Delta v_\varepsilon = f \quad \text{in } \Omega, \qquad \varepsilon \frac{\partial v_\varepsilon}{\partial \eta} + \beta v_\varepsilon = 0 \quad \text{in } \Gamma.$$

If we define the operators $A: L_2(\Omega) \to L_2(\Omega)$ and $A_\varepsilon : L_2(\Omega) \to L_2(\Omega)$ by

$$Av \equiv \Delta v, \qquad v \in \mathscr{D}(A) \equiv H^2(\Omega) \cap H_0^1(\Omega), \quad \text{and}$$

$$A_\varepsilon v_\varepsilon \equiv \Delta v_\varepsilon \qquad v_\varepsilon \in \mathscr{D}(A_\varepsilon) = \left\{ u \in L_2(\Omega);\ \Delta u \in L_2(\Omega);\ \varepsilon \frac{\partial u}{\partial \eta} + \beta u = 0 \right\},$$

then (3.1) and (3.2) are equivalent to

$$(3.1')\qquad Av = f, \quad \text{and}$$

$$(3.2')\qquad A_\varepsilon v_\varepsilon = f.$$

Below we shall state a number of regularity and convergence results established for the problems (3.1) and (3.2). The proofs of these results are relegated to § 3.2.

LEMMA 3.1.

$$(3.3)\qquad |A_\varepsilon^{-1} f|_{H^1(\Omega)} + \frac{1}{\sqrt{\varepsilon}} |A_\varepsilon^{-1} f|_{H^1(\Gamma)} \leqq C |f|_{H^{-1}(\Omega)},$$

$$(3.4)\qquad |A_\varepsilon^{-1} f|_{H^2(\Omega)} + \frac{1}{\varepsilon} |A_\varepsilon^{-1} f|_{H^2(\Gamma)} \leqq C \|f\|,$$

$$(3.5)\qquad |A_\varepsilon^{-1} f - A^{-1} f|_{H^2(\Omega)} \leqq C\varepsilon |f|_{H^{-1}(\Omega)}.$$

Next, let us define the so-called Dirichlet map $D: L_2(\Gamma) \to L_2(\Omega)$ *by*

$$\Delta Dg = 0 \quad \text{in } \Omega, \qquad Dg|_\Gamma = g \quad \text{in } \Gamma. \tag{3.6}$$

It is well known [LM] *that*

$$D \in \mathscr{L}(H^s(\Gamma) \to H^{s+1/2}(\Omega)) \quad \text{for all real } s > 0. \tag{3.7}$$

Similarly we define a "variational approximation" of (3.6) *by introducing the map* $N_\varepsilon: L_2(\Gamma) \to L_2(\Omega)$ *where*

$$\Delta N_\varepsilon g = 0 \quad \text{in} \quad \Omega, \qquad \varepsilon \frac{\partial N_\varepsilon}{\partial \eta} g + \beta N_\varepsilon g = \beta g \quad \text{in } \Gamma. \tag{3.8}$$

We shall prove the following lemma.

LEMMA 3.2.

$$|N_\varepsilon g|_{H^{3/2}(\Omega)} + \left| \frac{\partial}{\partial \eta} N_\varepsilon g \right| \leqq C|g|_{H^1(\Gamma)}, \tag{3.9}$$

$$|N_\varepsilon g|_{H^{1/2}(\Omega)} + \left| \frac{\partial}{\partial \eta} N_\varepsilon g \right|_{H^{-1}(\Gamma)} \leqq C|g|, \tag{3.10}$$

$$|N_\varepsilon g|_{H^2(\Omega)} \leqq C|g|_{H^{3/2}(\Gamma)}, \tag{3.11}$$

$$|N_\varepsilon g - Dg|_{H^{3/2}(\Omega)} \leqq C\varepsilon|g|, \tag{3.12}$$

$$|N_\varepsilon g - Dg|_{H^{5/2}(\Omega)} \leqq C\varepsilon|g|_{H^1(\Gamma)}. \tag{3.13}$$

Remark 3.1. Notice that the regularity properties (3.3)-(3.4), and (3.9)-(3.11), reconstruct (uniformly in the parameter $\varepsilon > 0$) the well-known regularity properties of the elliptic Dirichlet problems.

Remark 3.2. The results of Lemma 3.2 can be easily generalized to obtain $N_\varepsilon \in \mathscr{L}(H^s(\Gamma) \to H^{s+1/2}(\Omega))$ for all real $s > 0$ with the norm uniform in $\varepsilon > 0$.

3.2. Proofs of Lemma 3.1 and Lemma 3.2.

Proof of Lemma 3.1. With A_ε introduced as in § 3.1 we associate the bilinear form $a_\varepsilon(u, v)$ defined by

$$a_\varepsilon(u, v) \equiv -(A_\varepsilon u, v) = (\nabla u, \nabla v) + \frac{1}{\varepsilon}\langle \beta u, v \rangle$$

for all $u, v \in \mathscr{D}(A_\varepsilon^{1/2}) \equiv \{u \in H^1(\Omega), (1/\sqrt{\varepsilon})u|_\Gamma \in H^1(\Gamma)\}$.

It follows by standard arguments that $A_\varepsilon = A_\varepsilon^*$ and

$$\begin{aligned} &\text{(a)} \quad |a_\varepsilon(u, v)| \leqq C\left[|u|_{H^1(\Omega)}|v|_{H^1(\Omega)} + \frac{1}{\varepsilon}|u|_{H^1(\Gamma)}|v|_{H^1(\Gamma)} \right], \\ &\text{(b)} \quad |a_\varepsilon(u, u)| \geqq c\left[|u|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|u|^2_{H^1(\Gamma)} \right], \qquad c, C > 0. \end{aligned} \tag{3.14}$$

With the above notation, solving (3.2′) is equivalent to finding v_ε such that

$$a_\varepsilon(v_\varepsilon, \phi) = (f, \phi) \quad \text{for all } \phi \in \mathscr{D}(A_\varepsilon^{1/2}). \tag{3.15}$$

Property (3.3) now follows immediately from (3.14) after setting $\phi = v_\varepsilon$ in (3.15). As for (3.4) we first notice that

$$|v_\varepsilon|_{H^1(\Gamma)} \leqq C\sqrt{\varepsilon}|f|_{H^{-1}(\Omega)}$$

together with

$$\Delta v_\varepsilon = f \in L_2(\Omega) \tag{3.16}$$

imply via standard results from elliptic theory (see for example [B1])

$$|v_\varepsilon|_{H^{3/2}(\Omega)} + \left|\frac{\partial v_\varepsilon}{\partial \eta}\right|_{L_2(\Gamma)} \leqq C[\sqrt{\varepsilon}|f|_{H^{-1}(\Omega)} + |f|_{H^{-1/2}(\Omega)}]. \tag{3.17}$$

Thus (3.17) and $\varepsilon(\partial v_\varepsilon/\partial \eta) + \beta v_\varepsilon = 0$ yield

$$|v_\varepsilon|_{H^2(\Gamma)} \leqq C|\beta v_\varepsilon| \leqq C\varepsilon\left|\frac{\partial v_\varepsilon}{\partial \eta}\right| \leqq C\varepsilon[\sqrt{\varepsilon}|f|_{H^{-1}(\Omega)} + |f|_{H^{-1/2}(\Omega)}]. \tag{3.18}$$

Using (3.18), (3.16) and well-known energy estimates for the Dirichlet problem [LM], we obtain

$$|v_\varepsilon|_{H^2(\Omega)} \leqq C[|v_\varepsilon|_{H^{3/2}(\Gamma)} + \|f\|] \leqq C\|f\|. \tag{3.19}$$

Formulae (3.18) and (3.19) yield (3.4).

As for (3.5), we notice first that by using regularity results from elliptic theory applied to (3.1′) we obtain

$$\left|\frac{\partial v}{\partial \eta}\right|_{H^{-1/2}(\Gamma)} \leqq C|f|_{H^{-1}(\Omega)}.$$

Consequently,

$$\left|\beta^{-1}\frac{\partial v}{\partial \eta}\right|_{H^{3/2}(\Gamma)} \leqq C\left|\frac{\partial v}{\partial \eta}\right|_{H^{-1/2}(\Gamma)} \leqq C|f|_{H^{-1}(\Omega)}. \tag{3.20}$$

Next we define

$$z_\varepsilon \equiv A_\varepsilon^{-1}f - A^{-1}f = v_\varepsilon - v.$$

We have

$$\Delta z_\varepsilon = 0, \qquad z_\varepsilon|_\Gamma = -\varepsilon\beta^{-1}\frac{\partial v}{\partial \eta} = -\varepsilon\beta^{-1}\frac{\partial}{\partial \eta}A^{-1}f.$$

Since $z_\varepsilon = -\varepsilon D\beta^{-1}\partial v/\partial \eta$, (3.20) together with (3.7) gives

$$|z_\varepsilon|_{H^2(\Omega)} \leqq C\varepsilon\left|\beta^{-1}\frac{\partial v}{\partial \eta}\right|_{H^{3/2}(\Gamma)} \leqq C\varepsilon|f|_{H^{-1}(\Omega)}$$

which completes the proof of the Lemma 3.1. □

Proof of Lemma 3.2. Let $v_\varepsilon \equiv N_\varepsilon g$. We have

$$a_\varepsilon(v_\varepsilon, \phi) = \frac{1}{\varepsilon}\langle \beta g, \phi\rangle, \qquad \phi \in \mathscr{D}(A_\varepsilon^{1/2}).$$

Setting $\phi \equiv v_\varepsilon$ and using (3.14) yields

$$|v_\varepsilon|^2_{H^1(\Omega)} + \frac{1}{\varepsilon}|v_\varepsilon|^2_{H^1(\Gamma)} \leqq \frac{c}{\varepsilon}|g|^2_{H^1(\Gamma)}. \tag{3.21}$$

Hence in particular

$$|v_\varepsilon|_{H^1(\Gamma)} \leqq C|g|_{H^1(\Gamma)} \quad \text{where } \Delta v_\varepsilon = 0. \tag{3.22}$$

The standard estimates for the Dirichlet problem applied to (3.22) now yield (3.9). To prove (3.10) we use transposition. Notice first that $N_\varepsilon \in \mathscr{L}(L_2(\Gamma) \to H^{1/2}(\Omega))$ is equivalent to

$$N_\varepsilon^* \in \mathscr{L}(H^{-1/2}(\Omega) \to L_2(\Gamma)) \tag{3.23}$$

(with the norm independent on $\varepsilon > 0$). On the other hand, it can be easily verified that

$$N_\varepsilon^* A_\varepsilon u = \frac{\partial}{\partial \eta} u, \qquad u \in \mathcal{D}(A_\varepsilon). \tag{3.24}$$

Indeed, for $u \in \mathcal{D}(A_\varepsilon)$ we have

$$\langle N_\varepsilon^* A_\varepsilon u, g\rangle = (A_\varepsilon u, N_\varepsilon g) = (\Delta u, N_\varepsilon g) = \left\langle \frac{\partial u}{\partial \eta}, N_\varepsilon g \right\rangle - \left\langle u, \frac{\partial}{\partial \eta} N_\varepsilon g \right\rangle$$

$$= \left\langle \frac{\partial u}{\partial \eta}, N_\varepsilon g \right\rangle + \left\langle \frac{\partial u}{\partial \eta}, g - N_\varepsilon g \right\rangle = \left\langle \frac{\partial u}{\partial \eta}, g \right\rangle$$

which proves (3.24).

In view of (3.24), (3.23) will be proved as soon as we show that for any $v \in H^{-1/2}(\Omega)$

$$\left| \frac{\partial u}{\partial \eta} \right| \leqq C |v|_{H^{-1/2}(\Omega)} \quad \text{where } u = A_\varepsilon^{-1} v. \tag{3.25}$$

On the other hand, from (3.3) we have

$$|u|_{H^1(\Gamma)} \leqq C\sqrt{\varepsilon}\, |v|_{H^{-1}(\Omega)} \leqq C\sqrt{\varepsilon}\, |v|_{H^{-1/2}(\Omega)}. \tag{3.26}$$

Formula (3.26) combined with $\Delta u = v \in H^{-1/2}(\Omega)$ and the standard Dirichlet estimates yield

$$|u|_{H^{3/2}(\Omega)} + \left| \frac{\partial u}{\partial \eta} \right| \leqq C[|v|_{H^{-1/2}(\Omega)} + |u|_{H^1(\Gamma)}] \leqq C|v|_{H^{-1/2}(\Omega)},$$

which in particular implies (3.25); hence we have (3.23) and consequently

$$|N_\varepsilon g|_{H^{1/2}(\Omega)} \leqq c|g|.$$

To complete the proof of (3.10) we need to show that

$$\left| \frac{\partial}{\partial \eta} N_\varepsilon g \right|_{H^{-1}(\Gamma)} \leqq C|g|. \tag{3.27}$$

We first prove that for any $0 < \rho < 1$

$$\left| \frac{\partial}{\partial \eta} N_\varepsilon g \right|_{H^{-1-\rho}(\Gamma)} \leqq C|g|. \tag{3.28}$$

To see this, we apply the Green identity to $(N_\varepsilon g, \Delta\phi)$ with $\phi \in C^\infty(\Omega)$. This yields

$$\left\langle \frac{\partial}{\partial \eta} N_\varepsilon g, \phi \right\rangle = \left\langle N_\varepsilon g, \frac{\partial}{\partial \eta} \phi \right\rangle + (N_\varepsilon g, \Delta\phi), \qquad \phi \in C^\infty(\Omega).$$

The surjectivity of the trace operator (see [LM]) implies that for any $\tilde{\phi} \in H^{1+\rho}(\Gamma)$ we can select $\phi \in H^{3/2+\rho}(\Omega)$ such that

$$\frac{\partial}{\partial \eta} \phi|_\Gamma = 0 \quad \text{and} \quad \tilde{\phi}|_\Gamma = \phi_\Gamma.$$

Therefore

$$\left\langle \frac{\partial}{\partial \eta} N_\varepsilon g, \tilde{\phi} \right\rangle = (N_\varepsilon g, \Delta\phi) \quad \text{for any } \tilde{\phi} \in H^{1+\rho}(\Gamma) \text{ where } \phi \in H^{3/2+\rho}(\Omega). \tag{3.29}$$

Since $N_\varepsilon \in \mathscr{L}(L_2(\Gamma) \to H^{1/2}(\Omega))$ (by (3.23)) and $\Delta\phi \in H^{-1/2}(\Omega)$ for any $g \in L_2(\Gamma)$, $(\partial/\partial\eta)N_\varepsilon g$ defines a linear bounded functional on $H^{1+\rho}(\Gamma)$. Formula (3.28) now follows from (3.29). To prove (3.27) we write

$$\Delta N_\varepsilon g = 0, \qquad N_\varepsilon g|_\Gamma = g - \varepsilon\beta^{-1}\frac{\partial N_\varepsilon}{\partial\eta}g. \tag{3.30}$$

Since by the virtue of (3.28)

$$\left|\beta^{-1}\frac{\partial N_\varepsilon}{\partial\eta}g\right| \leqq C\left|\frac{\partial N_\varepsilon}{\partial\eta}g\right|_{H^{-2}(\Gamma)} \leqq C\left|\frac{\partial N_\varepsilon}{\partial\eta}g\right|_{H^{-1-\rho}(\Gamma)} \leqq C|g|,$$

the standard Dirichlet estimates applied to (3.30) yield

$$|N_\varepsilon g|_{H^{1/2}(\Omega)} + \left|\frac{\partial}{\partial\eta}N_\varepsilon g\right|_{H^{-1}(\Gamma)} \leqq C\left[|g| + \varepsilon\left|\beta^{-1}\frac{\partial N_\varepsilon}{\partial\eta}g\right|\right] \leqq C|g|,$$

which proves (3.10).

In order to establish (3.11) we return to (3.30). From (3.9) we have

$$\left|\frac{\partial}{\partial\eta}N_\varepsilon g\right| \leqq C|g|_{H^1(\Gamma)} \leqq C|g|_{H^{3/2}(\Gamma)}.$$

Thus

$$\left|\beta^{-1}\frac{\partial}{\partial\eta}N_\varepsilon g\right|_{H^{3/2}(\Gamma)} \leqq C\left|\frac{\partial}{\partial\eta}N_\varepsilon g\right| \leqq C|g|_{H^{3/2}(\Gamma)}.$$

Since $N_\varepsilon g|_\Gamma = g - \varepsilon\beta^{-1}(\partial/\partial\eta)N_\varepsilon g$,

$$|N_\varepsilon g|_{H^{3/2}(\Gamma)} \leqq C|g|_{H^{3/2}(\Gamma)}. \tag{3.31}$$

Formula (3.31) combined with $\Delta N_\varepsilon g = 0$ and standard Dirichlet estimates give (3.11). As for (3.12) we set

$$z \equiv (N_\varepsilon - D)g.$$

Then

$$\Delta z = 0, \qquad \varepsilon\frac{\partial}{\partial\eta}z + \beta z = \varepsilon\beta\beta^{-1}\frac{\partial}{\partial\eta}Dg. \tag{3.32}$$

Since

$$\left|\beta^{-1}\frac{\partial}{\partial\eta}Dg\right|_{H^1(\Gamma)} \leqq C\left|\frac{\partial}{\partial\eta}Dg\right|_{H^{-1}(\Gamma)} \leqq C|g| \tag{3.33}$$

and $z = \varepsilon N_\varepsilon(\beta^{-1}(\partial/\partial\eta)Dg)$, (3.12) immediately follows from (3.33) and (3.9). Finally, to prove (3.13) we notice that

$$z|_\Gamma = [N_\varepsilon g - Dg]|_\Gamma = -\varepsilon\beta^{-1}\frac{\partial N_\varepsilon}{\partial\eta}g.$$

Hence by (3.9)

$$|z|_{H^2(\Gamma)} \leqq C\varepsilon\left|\frac{\partial N_\varepsilon}{\partial\eta}g\right| \leqq C\varepsilon|g|_{H^1(\Gamma)}. \tag{3.34}$$

Formula (3.34) combined with $\Delta z = 0$ and the standard results for the Dirichlet problems yield (3.13), hence completing the proof of the Lemma 3.2. □

4. Differentiability of the solution to nonhomogeneous boundary problems—Proof of Theorem 2.3. Let u_ε be the solution to (1.2) with $f \equiv 0$, $g \in H^{1,1}(\Sigma)$, such that $g(0) = 0$. We first multiply (1.2) by $\dot{u}_\varepsilon$, integrate over Q and use the boundary conditions $u|_\Gamma = g - \varepsilon\beta^{-1}(\partial u/\partial\eta)$.

This leads to

$$\begin{aligned}(4.1)\qquad &\|\dot{u}_\varepsilon(T)\|^2 + \|\nabla u_\varepsilon(T)\|^2 + \varepsilon\left|\beta^{-1/2}\frac{\partial u_\varepsilon}{\partial\eta}(T)\right|^2 + 2\int_0^T \left\langle \frac{\partial u_\varepsilon}{\partial\eta}, g\right\rangle dt \\ &\leqq 2\rho\left|\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_2(\Sigma)} + \frac{1}{2\rho}|g|^2_{L_2(\Sigma)} \quad \text{for } \rho > 0.\end{aligned}$$

Let $\mathbf{h}(x) \subset C^1(\bar{\Omega})$ be such that $\mathbf{h}|_\Gamma = \mathbf{n}(x)$ where $\mathbf{n}$ is a normal vector on the boundary. We multiply (1.2) by $\nabla u\ \mathbf{h}$ and integrate over Q. This gives

$$\begin{aligned}&(\dot{u}_\varepsilon(T), h\nabla u_\varepsilon(T)) - \frac{1}{2}|\dot{u}_\varepsilon|^2_{L_2(\Sigma)} + \frac{1}{2}\int_Q ((u_\varepsilon)^2 - |\nabla u_\varepsilon|^2)\operatorname{div}\mathbf{h}\, dQ \\ &\quad = \left|\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_2(\Sigma)} - \frac{1}{2}\int_\Sigma |\nabla u_\varepsilon|^2 \mathbf{h}\mathbf{n}\, d\Sigma - \sum_{ij=1}^n \int_Q \frac{\partial u_\varepsilon}{\partial x_j}\frac{\partial h_i}{\partial x_j}\frac{\partial u_\varepsilon}{\partial x_i}\, dQ.\end{aligned}$$

Writing $\nabla u_\varepsilon|_\Gamma = \partial u_\varepsilon/\partial\eta + \nabla_\Gamma u_\varepsilon$ where

$$|\nabla_\Gamma u| \leqq C|u|_{H^1(\Gamma)},$$

we obtain

$$\left|\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_2(\Sigma)} + |\dot{u}_\varepsilon|^2_{L_2(\Sigma)} \leqq C[|\dot{u}_\varepsilon|^2_{L_\infty[0T;L_2(\Omega)]} + |u_\varepsilon|^2_{L_\infty[0T;H^1(\Omega)]} + |u_\varepsilon|^2_{L_2[0T;H^1(\Gamma)]}].$$

On the other hand, from

$$\beta^{1/2}u_\varepsilon = \beta^{1/2}g - \varepsilon\beta^{-1/2}\frac{\partial u_\varepsilon}{\partial\eta}$$

we have

$$(4.2)\qquad |u_\varepsilon|_{H^1(\Gamma)} \leqq C\left[|g|_{H^1(\Gamma)} + \varepsilon\left|\beta^{-1/2}\frac{\partial u_\varepsilon}{\partial\eta}\right|\right].$$

Consequently,

$$\begin{aligned}(4.3)\qquad &\left|\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_2(\Sigma)} + |\dot{u}_\varepsilon|^2_{L_2(\Sigma)} \\ &\leqq C\left[|\dot{u}_\varepsilon|^2_{L_\infty[0T;L_2(\Omega)]} + |u_\varepsilon|^2_{L_\infty[0T;H^1(\Omega)]} + |g|^2_{L_2[0T;H^1(\Gamma)]} + \varepsilon^2\left|\beta^{-1/2}\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_2(\Sigma)}\right].\end{aligned}$$

After combining the estimates in (4.1) and (4.3) and noticing that $|u_\varepsilon|_{H^1(\Omega)} \leqq C\|\nabla u_\varepsilon\|$, we obtain

$$\begin{aligned}(4.4)\qquad &|\dot{u}_\varepsilon|^2_{L_\infty[0T;L_2(\Omega)]} + |u_\varepsilon|^2_{L_\infty[0T;H^1(\Omega)]} + \varepsilon\left|\beta^{-1/2}\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_\infty[0T;L_2(\Gamma)]} \\ &\leqq 2\rho c\Big[|\dot{u}_\varepsilon|^2_{L_\infty[0T;L_2(\Omega)]} + |u_\varepsilon|^2_{L_\infty[0T;H^1(\Omega)]} + |g|^2_{L_2[0T;H^1(\Gamma)]} \\ &\qquad + \varepsilon^2\left|\beta^{-1/2}\frac{\partial u_\varepsilon}{\partial\eta}\right|^2_{L_2(\Sigma)}\Big] + \frac{C}{\rho}|g|_{L_2(\Sigma)}\end{aligned}$$

where the constant C does not depend on either ρ or $\varepsilon > 0$. Formula (2.9) now follows after taking in (4.4) ρ small enough (such that $2\rho C < 1$ and $2\rho C\varepsilon < 1$). As for (2.10), we first notice that (4.4) also implies that

$$\sqrt{\varepsilon}\left|\beta^{-1/2}\frac{\partial u}{\partial \eta}\right|_{L_\infty[0T;L_2(\Gamma)]} \leqq C|g|_{H^{1,1}(\Sigma)}. \tag{4.5}$$

Combining (2.9), (4.5), and (4.3) yields

$$\left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)} + |u_\varepsilon|_{L_2(\Sigma)} \leqq C|g|_{H^{1,1}(\Sigma)}. \tag{4.6}$$

Going back to (4.2) and making use of (4.6) provides us with the estimate

$$|u_\varepsilon|_{L_2[0T;H^1(\Gamma)]} \leqq C\left[|g|_{L_2[0T;H^1(\Gamma)]} + \left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)}\right] \leqq C|g|_{H^{1,1}(\Sigma)}. \tag{4.7}$$

Formulae (4.6) and (4.7) yield the desired result stated in (2.10).

Finally we shall prove (2.11). Indeed, since $\beta[u_\varepsilon - g]|_\Gamma = \varepsilon(\partial u\varepsilon/\partial \eta)$, (2.11) follows from

$$|u_\varepsilon - g|_{L_2[0T;H^2(\Gamma)]} \leqq C|\beta(u_\varepsilon - g)|_{L_2(\Sigma)} \leqq C\varepsilon\left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C\varepsilon|g|_{H^{1,1}(\Sigma)},$$

where in the last inequality again we have used (4.6).

Next we prove part (ii) of the theorem for $s = 2$. Set $p \equiv \dot{u}_\varepsilon$ where u_ε satisfies (1.2) with $g \in H^{2,2}(\Sigma)$. Then

$$\ddot{p} = \Delta p, \quad p(0) = \dot{p}(0) = 0, \quad \varepsilon\frac{\partial p}{\partial \eta} + \beta p = \beta \dot{g}. \tag{4.8}$$

Formulae (2.9), (2.10) applied to (4.8) yield

$$|\dot{u}_\varepsilon|_{C[0T;H^1(\Omega)]} + |\ddot{u}_\varepsilon|_{C[0T;L_2(\Omega)]} + |\dot{u}_\varepsilon|_{H^{1,1}(\Sigma)} + \left|\frac{\partial \dot{u}_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C|g|_{H^{2,2}(\Sigma)}. \tag{4.9}$$

Therefore we have a situation where

$$\Delta u_\varepsilon = \ddot{u}_\varepsilon \in C[0T;\, L_2(\Omega)], \qquad \varepsilon\frac{\partial u_\varepsilon}{\partial \eta} + \beta u_\varepsilon = \beta g, \quad g \in H^{2,2}(\Sigma).$$

We write

$$u_\varepsilon = A_\varepsilon^{-1}(\ddot{u}_\varepsilon) + N_\varepsilon g. \tag{4.10}$$

By the virtue of (3.4) and (3.11) applied to (4.10) we obtain

$$|u_\varepsilon|_{L_\infty[0T;H^2(\Omega)]} \leqq C[|\ddot{u}_\varepsilon|_{L_\infty[0T;L_2(\Omega)]} + |g|_{L_\infty[0T;H^{3/2}(\Gamma)]}]. \tag{4.11}$$

On the other hand, we have

$$|g|_{L_\infty[0T;H^{3/2}(\Gamma)]} \leqq C|g|_{H^{2,2}(\Sigma)} \quad \text{(see [LM, Thm. 3.1, p. 19])}. \tag{4.12}$$

Combining (4.9), (4.11), and (4.12) yields

$$|u_\varepsilon|_{L_\infty[0T;H^2(\Omega)]} \leqq C|g|_{H^{2,2}(\Sigma)}, \tag{4.13}$$

which together with (4.9) completes the proof of (2.12) with $s = 2$.

To show (2.13), in view of (4.9), it is enough to establish the following estimate:

$$\left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{L_2[0T;H^1(\Omega)]} + |u_\varepsilon|_{L_2[0T;H^2(\Gamma)]} \leqq C|g|_{H^{2,2}(\Sigma)}. \tag{4.14}$$

To accomplish this, let us introduce the operator $T=\sum_{ij=1}^{n} t_i(x)(\partial/\partial x_i)$ where the coefficients $t_i(x)$ are smooth on $\bar{\Omega}$ and T is tangential to the boundary Γ (i.e., $\sum_{i=1}^{n} t_i(x)n_i(x)=0$ $x\in\Gamma$). Proving (4.14) is equivalent to showing that

$$\left|\frac{\partial}{\partial\eta}Tu_\varepsilon\right|_{L_2(\Sigma)}+|Tu_\varepsilon|_{L_2[0T;H^1(\Gamma)]}\leqq C|g|_{H^{1,1}(\Sigma)}. \tag{4.15}$$

Setting $p\equiv Tu_\varepsilon$ gives

$$\begin{aligned}&\text{(i)}\quad \ddot{p}=\Delta p+[\Delta, T]u_\varepsilon,\\ &\text{(ii)}\quad p(0)=\dot{p}(0)=0,\\ &\text{(iii)}\quad \varepsilon\frac{\partial p}{\partial\eta}+\beta p=\beta\beta^{-1}\left[[T,\beta]u_\varepsilon+\varepsilon\left[T,\frac{\partial}{\partial\eta}\right]u_\varepsilon+T\beta g\right]\end{aligned} \tag{4.16}$$

where $[A, B]$ stands for the commutator between operators A and B. Since T is a first order differential operator, we have

$$|[\Delta, T]u_\varepsilon|_{L_1[0T;L_2(\Omega)]}\leqq C|u_\varepsilon|_{L_1[0T;H^2(\Omega)]}\leqq C|g|_{H^{2,2}(\Sigma)} \tag{4.17}$$

where in the last inequality we have used (4.13).

Next we shall estimate all three terms on the right-hand side of (4.16)(iii).

$$\begin{aligned}|\beta^{-1}[T,\beta]u_\varepsilon|_{H^{1,1}(\Sigma)}&\leqq C[|[T,\beta]u_\varepsilon|_{L_2[0T;H^{-1}(\Gamma)]}+|[T,\beta]u_\varepsilon|_{L_2[0T;H^{-2}(\Gamma)]}]\\ &\leqq C[|u_\varepsilon|_{L_2[0T;H^1(\Gamma)]}+|u_\varepsilon|_{L_2(\Sigma)}]\\ &\leqq C|g|_{H^{1,1}(\Sigma)}\end{aligned} \tag{4.18}$$

where in the last inequality we have used (2.10)

$$\begin{aligned}\left|\beta^{-1}\left[T,\frac{\partial}{\partial\eta}\right]u_\varepsilon\right|_{H^{1,1}(\Sigma)}&\leqq C\left|\left[T,\frac{\partial}{\partial\eta}\right]u_\varepsilon\right|_{L_2[0T;H^{-1}(\Gamma)]}+\left|\left[T,\frac{\partial}{\partial\eta}\right]u_\varepsilon\right|_{L_2[0T;H^2(\Gamma)]}\\ &\leqq C\left[\left|\frac{\partial}{\partial\eta}u_\varepsilon\right|_{L_2(\Sigma)}+|u_\varepsilon|_{L_2(\Sigma)}+\left|\frac{\partial}{\partial\eta}u_\varepsilon\right|_{L_2(\Sigma)}+|u_\varepsilon|_{L_2(\Sigma)}\right]\\ &\leqq C|g|_{H^{2,2}(\Sigma)}.\end{aligned} \tag{4.19}$$

$$|\beta^{-1}T\beta g|_{H^{1,1}(\Sigma)}\leqq C|\beta^{1/2}g|_{H^{1,1}(\Sigma)}\leqq C|g|_{H^{2,2}(\Sigma)}. \tag{4.20}$$

Formula (4.16) can now be rewritten as

$$\ddot{p}=\Delta p+F,\quad p(0)=\dot{p}(0)=0,\quad \varepsilon\frac{\partial p}{\partial\eta}+\beta p=\beta G \tag{4.21}$$

where by (4.17)-(4.20) we have

$$|F|_{L_1[0T;L_2(\Omega)]}\leqq C|g|_{H^{2,2}(\Sigma)},\qquad |G|_{H^{1,1}(\Sigma)}\leqq C|g|_{H^{2,2}(\Sigma)}.$$

Thus we are in a position to apply the estimates (2.10), (2.4),and (2.5) to (4.21). This yields

$$\left|\frac{\partial p}{\partial\eta}\right|_{L_2(\Sigma)}+|p|_{L_2[0T;H^1(\Gamma)]}\leqq C|g|_{H^{2,2}(\Sigma)}$$

which is precisely (4.15). Thus the proof of (2.13) is completed. As for (2.14), we simply write

$$u_\varepsilon-g=\varepsilon\beta^{-1}\frac{\partial u}{\partial\eta}.$$

Formula (2.14) now follows from (2.13). Thus we have proved (2.12)-(2.14) for $s=2$. The proof of these statements for all integer values of s greater than two follows along the same line as for $s=2$. The results for $s \geqq 1$, s-real, can be easily obtained via interpolation. □

5. Proof of Theorem 2.4. Let u_ε be the solution to (1.2) with $g=0$. From (2.4) in particular we have

$$\left|\frac{\partial u_\varepsilon}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C|f|_{L_1[0T;L_2(\Omega)]}.$$

Formula (2.15)(i) now follows after noticing that $u_\varepsilon|_\Gamma = -\varepsilon\beta^{-1}(\partial u/\partial\eta)$ and

$$|\beta^{-1}u|_{H^2(\Gamma)} \leqq C|u|.$$

As for (2.15)(ii), we set

$$z \equiv u_\varepsilon - u \text{ where } u \text{ satisfies (1.1) with } g=0.$$

Then

$$\ddot{z} = \Delta z, \quad z(0)=\dot{z}(0)=0, \quad \varepsilon\frac{\partial z}{\partial\eta}+\beta z = -\varepsilon\frac{\partial u}{\partial\eta}. \tag{5.1}$$

Formula (2.7) applied to (5.1) yields

$$|z|_{C[0T;L_2(\Omega)]} \leqq C\varepsilon\left|\frac{\partial u}{\partial\eta}\right|_{L_2(\Sigma)} \leqq C\varepsilon|f|_{L_1[0T;L_2(\Omega)]}$$

where in the last inequality we have used (2.2).

To establish (2.16) we shall need the following lemma.

LEMMA 5.1. *Let u_ε be the solution to* (1.2) *with $g\equiv 0$. Then for any $f\in L_1[0T;\ L_2(\Omega)]$ we have $\dot{u}_\varepsilon|_\Gamma \to 0$ in $L_2(\Sigma)$.*

Proof. To prove the lemma, in view of (2.4), it is enough to show that for f in some dense set $\mathcal{F}\in L_1[0T;\ L_2(\Omega)]$ we have

$$\dot{u}_\varepsilon \to 0 \quad \text{in } L_2(\Sigma). \tag{5.2}$$

To this end let $\mathcal{F}\equiv\{f\in H^1[0T;\ L_2(\Omega)],\ f(0)=0\}$, dense set in $L_1[0T;\ L_2(\Omega)]$.

If u_ε satisfies (1.2) with $g\equiv 0$ and the right-hand side of the equation is equal to f, then $p=\dot{u}_\varepsilon$ and $\tilde{f}\equiv\dot{f}$ satisfy

$$\ddot{p} = \Delta p + \tilde{f}, \quad p(0)=\dot{p}(0)=0, \quad \varepsilon\frac{\partial p}{\partial\eta}+\beta p = 0. \tag{5.3}$$

Energy estimate (2.5) applied to (5.3) yields

$$|p|_{C[0T;H^1(\Gamma)]} \leqq C\sqrt{\varepsilon}|\tilde{f}|_{L_1[0T;L_2(\Omega)]} \leqq C\sqrt{\varepsilon}|f|_{H^1[0T;L_2(\Omega)]}.$$

This proves that for $f\in\mathcal{F}$, $\dot{u}_\varepsilon\to 0$ in $C[0T;\ H^1(\Gamma)]$, hence in particular (5.2). □

To continue with the proof of Theorem 2.4 let us define $z\equiv u_\varepsilon - u$ where u_ε (respectively, u) satisfy (1.2) (respectively, (1.1)) with $g\equiv 0$. Then

$$\ddot{z} = \Delta z, \quad z(0)=z(0)=0, \quad \varepsilon\frac{\partial}{\partial\eta}z+\beta z = -\varepsilon\frac{\partial u}{\partial\eta}. \tag{5.4}$$

Multiplying (5.4) by $\dot{z}$ and integrating over Q leads to

$$\begin{aligned} &|\dot{z}|^2_{L_\infty[0T;L_2(\Omega)]}+|z|^2_{L_\infty[0T;H^1(\Omega)]}+\frac{1}{\varepsilon}|z|^2_{L_\infty[0T;H^1(\Gamma)]} \\ &\leqq 2\left|\frac{\partial u}{\partial\eta}\right|_{L_2(\Sigma)}|\dot{z}|_{L_2(\Sigma)} = 2\left|\frac{\partial u}{\partial\eta}\right|_{L_2(\Sigma)}|\dot{u}_\varepsilon|_{L_2(\Sigma)}. \end{aligned} \tag{5.5}$$

Application of Lemma 5.1 to (5.5), combined with the usual density argument, leads to the conclusion (2.16).

Statement (2.17) (i) is the same as that of Lemma 5.1. To prove (2.17)(ii) we argue as in the step (ii) of the proof of Theorem 2.3. In fact, multiplying (5.4) by $\Delta z\,\mathbf{h}$, integrating over Q and using the same arguments as step (ii) of Theorem 2.3 leads to

$$\left|\frac{\partial z}{\partial \eta}\right|^2_{L_2(\Sigma)} + |z|^2_{L_2(\Sigma)} \leqq C[|z|^2_{L_\infty[0T;H^1(\Omega)]} + |\dot z|^2_{L_\infty[0T;L_2(\Omega)]} + |z|^2_{L_2[0T;H^1(\Gamma)]}]. \tag{5.6}$$

Formula (2.17)(ii) now follows from (2.16) and (2.15)(i) applied to (5.6) (since $z|\Gamma = u^\varepsilon|\Gamma$). To complete the proof of the theorem we need to establish (2.18) and (2.19).

Proof of (2.18). Let $z \equiv u_\varepsilon - u$ where u_ε (respectively, u) satisfies (1.2) (respectively, (1.1)) with $f = 0$ and $g \in H^{1,1}(\Sigma)$. Then

$$\ddot z = \Delta z, \quad z(0) = \dot z(0) = 0, \quad \varepsilon\frac{\partial z}{\partial \eta} + \beta z = -\varepsilon\frac{\partial u}{\partial \eta} = -\varepsilon\beta\beta^{-1}\frac{\partial u}{\partial \eta}. \tag{5.7}$$

Since by (2.2)

$$\left|\beta^{-1}\frac{\partial u}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C\left|\frac{\partial u}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C|g|_{H^{1,1}(\Sigma)},$$

the application of formula (2.7) of Theorem 2.2 gives

$$|z|_{C[0T;L_2(\Omega)]} \leqq C\varepsilon\left|\beta^{-1}\frac{\partial u}{\partial \eta}\right|_{L_2(\Sigma)} \leqq C\varepsilon|g|_{H^{1,1}(\Sigma)}$$

which is precisely the statement (2.18). Formula (2.19) can be deduced from (2.18) and (2.7) followed by the standard density argument. The proof of Theorem 2.4 is thus completed. □

REFERENCES

[L1] J. L. LIONS, *A remark on the approximation of nonhomogeneous boundary value problems*, in Vistas in Applied Mathematics, A. V. Balakrishnan, A. A. Dorodnicen, and J. L. Lions, eds., Optimation Software Inc., New York, 1986.

[L2] ———, *Controle des systèmes distribues singuliers*, Dunod, Paris, 1983.

[B1] I. BABUSKA AND A. AZIZ, *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, London, 1972.

[LM] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Vols. I, II, Springer-Verlag, Berlin, New York, Heidelberg, 1972.

[LLT] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Nonhomogenous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., to appear.

[LS] I. LASIECKA AND J. SOKOLOWSKI, *Semidiscrete approximations of hyperbolic boundary value problems with nonhomogenous Dirichlet boundary conditions*, submitted.

[LT1] I. LASIECKA AND R. TRIGGIANI, *A cosine operator approach to modelling $L_2(0T; L_2(\Gamma))$ boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35-83.

[LT2] ———, *Regularity of hyperbolic equations under $L_2(0T; L_2(\Gamma))$ Dirichlet boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275-286.

[S1] R. SAKOMOTO, *Hyperbolic Boundary Value Problems*, Cambridge University Press, Cambridge, London, 1982.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

004

THE SOLUTION OF THE RIEMANN PROBLEM FOR A HYPERBOLIC SYSTEM OF CONSERVATION LAWS MODELING POLYMER FLOODING*

THORMOD JOHANSEN† AND RAGNAR WINTHER‡

Abstract. The global Riemann problem for a nonstrictly hyperbolic system of conservation laws modeling polymer flooding is solved. In particular, the system contains a term that models adsorption effects.

Key words. Riemann problem, polymer flooding, adsorption

AMS(MOS) subject classifications. 35L65, 76S05

1. Introduction. In this paper we solve the global Riemann problem for the mathematical model

$$\begin{aligned} &s_t + f(s, c)_x = 0, \\ &[sc + a(c)]_t + [f(s, c)c]_x = 0, \end{aligned} \tag{1.1}$$

where $t \in \mathbf{R}^+$, $x \in \mathbf{R}$, the state vector $(s, c) \in I \times I$ and $f: I \times I \to \mathbf{R}$ and $a: I \to \mathbf{R}$ are smooth functions. Here I denotes the unit interval $I = [0, 1]$. More precise assumptions on the functions f and a will be given in § 2.

Our results generalize the results of Isaacson [6], where the Riemann problem for (1.1) is solved with the term $a(c)$ neglected. In this simplified case the solution of the Riemann problem can also be derived from the analysis given by Keyfitz and Kranzer [7].

We note that when c is constant, (1.1) reduces to the single equation

$$s_t + f(s)_x = 0, \tag{1.2}$$

which in the petroleum literature is known as the Buckley–Leverett equation.

The model (1.1) arises in connection with enhanced oil recovery, for example when oil is displaced in a porous rock by water containing dissolved polymer. The variable s is the saturation of the mixture of water and polymer, which we call the aqueous phase. The variable c is the concentration of polymer in the aqueous phase. Furthermore, f describes the fractional flow of the aqueous phase, which is assumed to be immiscible with oil. The function $a(c)$ models adsorption of the polymer on rock.

Some other processes governed by (1.1) are discussed by Pope in [9], where detailed physical assumptions for the model are also discussed.

The derivation of (1.1) is based on material balance considerations. We assume that the fluids and the rock are incompressible, and that volumes do not change when polymer is dissolved in water. Assuming one-dimensional flow in a homogeneous medium, the mass conservation of water, oil, and polymer, respectively, can be formulated as follows:

$$\phi s_t + v_x = 0, \tag{1.3a}$$

$$\phi s_t^0 + v_x^0 = 0, \tag{1.3b}$$

$$\phi[sc + a(c)]_t + [vc]_x = 0, \tag{1.3c}$$

* Received by the editors August 18, 1986; accepted for publication (in revised form) April 27, 1987. This work was supported by Statoil and The Norwegian Academy of Science through the VISTA-programme.

† Institute for Energy Technology, 2007 Kjeller, Norway.

‡ Institute of Informatics, The University of Oslo, Oslo 3, Norway.

where ϕ is the rock porosity, s^0 is the oil saturation and v and v^0 denote the volumetric flow rates of the aqueous phase and oil, respectively. Neglecting gravity, capillarity and dispersion, we get v and v^0 by Darcy's law as follows:

$$v = -\lambda p_x, \qquad \lambda = -\frac{Kk}{\mu}, \tag{1.4a}$$

$$v^0 = -\lambda^0 p_x, \qquad \lambda^0 = -\frac{Kk^0}{\mu^0}. \tag{1.4b}$$

Here, p is the fluid pressure, K is the absolute permeability of the rock, $k = k(s, c)$ and $k^0 = k^0(s)$ are the relative permeabilities of the aqueous phase and oil, respectively, and μ, μ^0 are the corresponding viscosities.

Since $s^0 + s \equiv 1$, summation of (1.3a) and (1.3b) gives

$$v^T \equiv v^0 + v = \text{constant} \tag{1.5}$$

where the value of total volumetric flow rate v^T is determined from the boundary conditions. By elimination of the pressure p using (1.4a), (1.4b), and (1.5), the equations (1.3a) and (1.3b) can be rewritten in the form

$$\phi s_t + v^T f_x = 0, \tag{1.6a}$$

$$\phi[sc + a(c)]_t + v^T[fc]_x = 0, \tag{1.6b}$$

which together with (1.5) constitutes the model. Here $f = f(s, c)$ denotes the fractional flow function given by

$$f = \frac{\lambda}{\lambda + \lambda^0}. \tag{1.7}$$

The dependence on s is inherited from the relative permeability functions k and k^0 while the dependence on c is primarily introduced through the viscosity $\mu = \mu(c)$ of the aqueous phase.

If L denotes the length of the medium, a simple coordinate transformation

$$x' = \frac{x}{L} \quad \text{and} \quad t' = \frac{v^T t}{\phi L}$$

results in (1.1), when the primes on x', t' are dropped.

The model (1.1) is an example of a system of hyperbolic conservation laws. The main result of this paper is the construction of a unique global solution of the Riemann problem for the system (1.1); i.e., we construct a weak solution of the pure initial value problem for (1.1) with initial condition

$$(s, c)(x, 0) = \begin{cases} (s^L, c^L) & \text{if } x < 0, \\ (s^R, c^R) & \text{if } x > 0, \end{cases} \tag{1.8}$$

where the left and right states (s^L, c^L) and (s^R, c^R) in $I \times I$ are arbitrary. In order to distinguish the physically meaningful weak solutions we will also require that the solution satisfies an "entropy condition." This condition will be derived by demanding that the solution is evolutionary; i.e., any discontinuity of the solution is a limit of smooth solutions of associated "viscosity systems." The solution of the Riemann problem in general depends only on the ratio x/t and it will be constructed by connecting constant states, smooth solutions (or rarefaction waves) and discontinuous solutions (or shock waves).

In addition to being analytical solutions of the system, solutions of the Riemann problem can also be used as building blocks for the construction of numerical methods for (1.1). Examples of this are the Random Choice Method [1], [2], [4], and Godunov's method [5]. Unfortunately, for many hyperbolic systems no global solution of the Riemann problem has yet been found. A general theory for local existence and uniqueness of solution is described in [4], [8], or [11] under the condition of strict hyperbolicity of the system. However, this condition will not be satisfied in the analysis of (1.1) given below.

In the case of the single Buckley–Leverett equation (1.2) (i.e., when c is constant) the solution of the Riemann problem is well known (cf. [3], [8] and § 4 below for a precise formulation of the entropy condition in this case). Assume for example that $s^L < s^R$ and let $g(s) = (d/ds)f_L(s)$, where f_L is the lower convex envelope of f with respect to the interval $[s^L, s^R]$. The unique solution of the Riemann problem is then given by

$$s(x,t) = \begin{cases} s^L & \text{if } x/t < g(s^L), \\ s & \text{if } x/t = g(s), \\ s^R & \text{if } x/t > g(s^R), \end{cases} \tag{1.9}$$

where we adopt the convention that if $g(s) = \sigma$ on a maximal interval (s^1, s^2), with $s^1 < s^2$, then this corresponds to a discontinuity at $x = \sigma t$ with $s(\sigma t-, t) = s^1$ and $s(\sigma t+, t) = s^2$. Similarly, if $s^R < s^L$ the unique solution of the Riemann problem is again given by (1.9), where in this case $g(s) = (d/ds)f_U(s)$ is the derivative of the upper concave envelope f_U of f with respect to $[s^R, s^L]$. The solution of the Riemann problem for (1.2) given above will be a fundamental part of the solution of the Riemann problem for the system (1.1) constructed in this paper.

The main difference between the present paper and [6] is that the adsorption term $a(c)$ has been included in the model here. The effect of this term is that the linearly degenerate characteristic field appearing in the analysis in [6] (and [7]) is replaced by a nondegenerate field. The contact discontinuities in the Riemann solution given in [6] will therefore be replaced by either proper rarefaction waves or proper shock waves. As a consequence, both the state space solution and the x, t-space solution of the Riemann problem are unique. This is in contrast to [6], where the solution is unique in x, t-space, but not in state space. Also, in [6] the Lax entropy inequalities were used as the entropy condition for the system. In this paper we derive entropy conditions from traveling wave analysis and we conclude that, in general, the Lax inequalities are not the correct entropy condition for the system. In particular, we derive an admissible shock wave where both characteristics on both sides of the shock enter the shock. Such shock waves are referred to as overcompressive shocks by Schaeffer and Shearer [10]. However, this wave cannot be joined to any other wave in a Riemann solution.

The construction of the solution of the Riemann problem for the model (1.1) given below shows that if the solution is considered pointwise, it is discontinuous with respect to the left and right states for certain critical values of the left and right states. However, the solution is continuous in L^1 norm (cf. Example 8.2). This property is similar to properties of Riemann solutions of the nonstrictly hyperbolic systems studied in, e.g., [6], [7], and [10].

The precise assumptions on the system (1.1) are stated in § 2. The simple rarefaction waves are derived in § 3, while § 4 is devoted to shock waves and entropy conditions. In § 5 we formulate the general Riemann problem in terms of rarefaction waves and shock waves and state the main result of the paper. The proof is given in §§ 6 and 7.

In § 8 we present some numerical experiments. One of the purposes of these experiments is to show that the behavior of the exact solution of the Riemann problem is not easily detected from calculations done by standard finite difference schemes.

2. A precise formulation of the model. We recall that our model for the polymer process is the following 2×2 system of conservation laws:

$$\begin{aligned} &s_t + f(s, c)_x = 0, \\ &(sc + a(c))_t + (cf(s, c))_x = 0, \end{aligned} \tag{2.1}$$

where the unknown functions are $s = s(x, t)$ and $c = c(x, t)$. Throughout the paper we will assume that the real-valued function $f = f(s, c)$ is a smooth function for $(s, c) \in I \times I$, where $I = [0, 1]$, with the following properties (cf. Fig. 2.1):

(i) $f(0, c) \equiv 0$, $f(1, c) \equiv 1$;

(ii) $f_s(s, c) > 0$ for $0 < s < 1$, $0 \leqq c \leqq 1$;

(iii) $f_c(s, c) < 0$ for $0 < s < 1$, $0 \leqq c \leqq 1$;

(iv) For each $c \in I$, $f(\cdot, c)$ has a unique point of inflection, $s^I = s^I(c) \in (0, 1)$, such that $f_{ss}(s, c) > 0$ for $0 < s < s^I$ and $f_{ss}(s, c) < 0$ for $s^I < s < 1$.

The function $a = a(c)$, which models the adsorption effects of the process, is assumed to be a smooth function of $c \in I$ such that (cf. Fig. 2.2):

(i) $a(0) = 0$;

(ii) $h(c) = (da/dc)(c) > 0$ for $0 < c < 1$;

(iii) $(dh/dc)(c) = (d^2a/dc^2)(c) < 0$ for $0 < c < 1$.

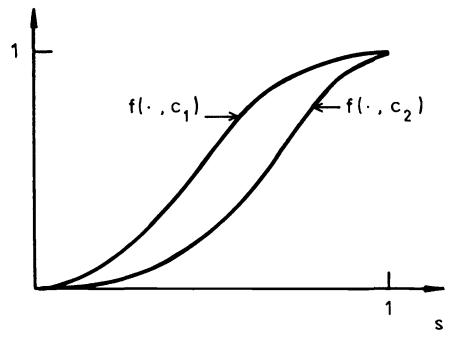


FIG. 2.1. $c_1 < c_2$.

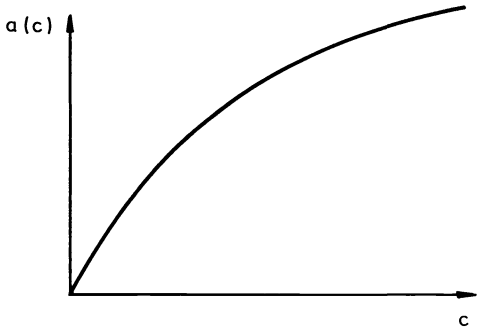


FIG. 2.2

Let u denote the state vector $u=(s, c)$. The system (2.1) may be reformulated in the form

$$u_t + A(u)u_x = 0,$$

where $A(u)=A(s, c)$ is the upper triangular 2×2 matrix

$$A(s, c)=\begin{bmatrix} f_s(s, c) & f_c(s, c) \\ 0 & \dfrac{f(s, c)}{s+h(c)} \end{bmatrix}. \tag{2.2}$$

The eigenvalues of A are $\lambda^s = f_s$ and $\lambda^c = f/(s+h)$, with corresponding eigenvectors $e^s=(1, 0)$, $e^c=(f_c, \lambda^c-\lambda^s)$ if $0<s<1$ and $e^c=(0, 1)$ if $s=0, 1$.

We observe (cf. Figs. 2.3, 2.4) that for each $c\in I$ there is at most one $s^T=s^T(c)\in I$ such that

$$\lambda^c(s^T, c)=\lambda^s(s^T, c). \tag{2.3}$$

Throughout the paper we will assume that there exists a unique $c^T\in I$ such that (2.3) has a unique solution $s^T(c)$ for $0\leqq c\leqq c^T$ and that (2.3) has no solution for $c^T<c\leqq 1$ (cf. Figs. 2.5, 2.6). We let T denote the transition curve

$$T=\{(s, c)\,|\,0\leqq c\leqq c^T,\ s=s^T(c)\}$$

and $\mathscr{L}$ and $\mathscr{R}$ the regions

$$\mathscr{L}=\{(s, c)\in I\times I\,|\,\lambda^s>\lambda^c\},\qquad \mathscr{R}=\{(s, c)\in I\times I\,|\,\lambda^c>\lambda^s\}.$$

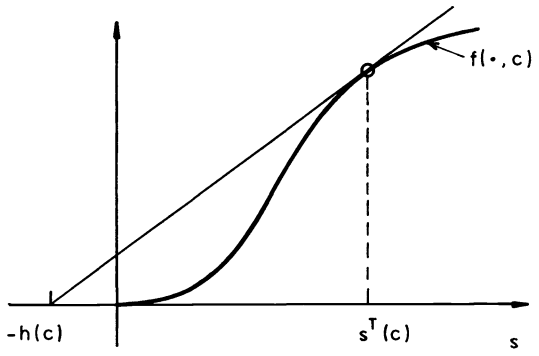


FIG. 2.3. $s^T(c)$ *exists.*

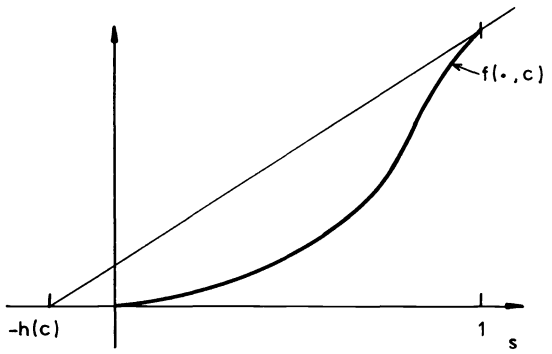


FIG. 2.4. *No* $s^T(c)$.

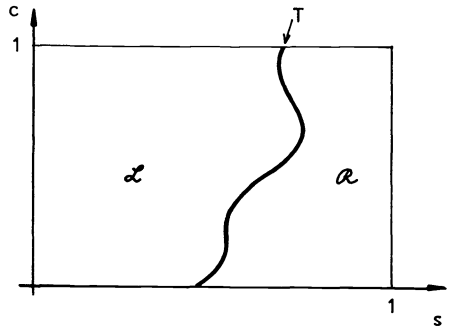


FIG. 2.5. $c^T = 1$.

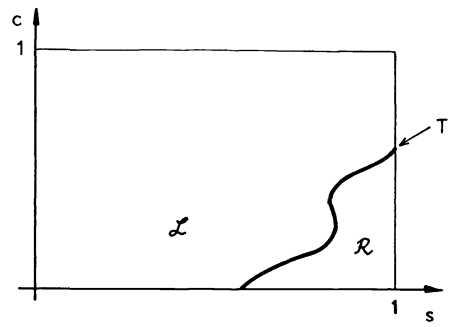


FIG. 2.6. $c^T < 1$.

We observe that if $(s, c) \in T$ and $0 < s < 1$, then $\lambda^s = \lambda^c$ and the two eigenvectors e^s and e^c become linearly dependent. Hence, in this case the matrix A given by (2.2) is not diagonalizable.

3. Rarefaction waves. The purpose of this section is to determine the simple rarefaction waves of the system (1.1). Hence, for two given states $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$ we derive possible continuous solutions of the pure initial value problem for (1.1) with initial data

$$u(x, 0) = \begin{cases} u^L & \text{if } x < 0, \\ u^R & \text{if } x > 0. \end{cases} \tag{3.1}$$

Let λ be an eigenvalue of the matrix A given by (2.2) with corresponding eigenvector e. The simple rarefaction waves are continuous solutions of (1.1) and (3.1) of the form

$$u(x, t) = v(x/t),$$

where v corresponds to an integral curve of the vector field e. More precisely,

$$u(x, t) = \begin{cases} u^L & \text{if } x/t < \lambda(u^L), \\ v & \text{if } x/t = \lambda(v), \\ u^R & \text{if } x/t > \lambda(u^R), \end{cases} \tag{3.2}$$

where v is an integral curve of the vector field e connecting the states u^L and u^R with the additional property that the eigenvalue λ is increasing from u^L to u^R. Since the

matrix A has two eigenvalues, λ^s and λ^c, there are two possible rarefaction curves through any given state u^L.

s-rarefaction waves. If $\lambda = \lambda^s$ and $e = e^s = (1, 0)$ the integral curves of e are the curves where c is constant. Hence, a simple rarefaction wave of the form (3.2) exists if $c^L = c^R$ and if $\lambda^s = f_s(s, c^L)$ is increasing from s^L to s^R. This, of course, corresponds to a simple rarefaction wave of the Buckley–Leverett equation (1.2) with $f(s) = f(s, c^L)$. Such rarefaction waves will be referred to as s-rarefaction waves.

c-rarefaction waves. Next we consider the case when $\lambda = \lambda^c$. First assume that $0 < s < 1$. In this case $f_c < 0$, the eigenvector e^c can be taken to be $(f_c, \lambda^c - \lambda^s)$, and the rarefaction curves are determined by the differential equation

$$f_c \frac{dc}{ds} = \lambda^c - \lambda^s, \tag{3.3}$$

where $c = c(s)$. Hence, $dc/ds > 0$ in region $\mathscr{L}$, $dc/ds < 0$ in $\mathscr{R}$ and $dc/ds = 0$ on the transition curve T. Furthermore, by differentiating (3.3) with respect to s we also obtain

$$f_c \frac{d^2c}{ds^2} = -f_{ss}$$

when $(s, c(s)) \in T$. Hence, $d^2c/ds^2 < 0$ on T. As a consequence, any rarefaction curve can at most intersect the transition curve T in one point (cf. Fig. 3.1).

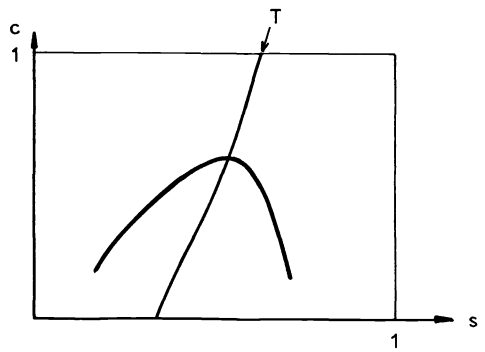


FIG. 3.1. *Rarefaction curve.*

Next consider $g(s) = \lambda^c(s, c(s))$, where $c(s)$ satisfies (3.3). A straightforward differentiation shows that

$$\frac{dg}{ds} = \frac{(f_s + f_c(dc/ds))(s+h) - f(1 + (dh/dc)(dc/ds))}{(s+h)^2}$$

or

$$\frac{dg}{ds}(s) = -\frac{(dh/dc)(c(s))(dc/ds)(s)}{(s + h(c(s)))^2} f(s, c(s)).$$

Since $(dh/dc)(c) < 0$ the eigenvalue $\lambda^c(s, c(s))$ is an increasing function of s in $\mathscr{L}$ and decreasing in $\mathscr{R}$. Hence, a simple rarefaction wave of the form (3.2) exists if and only if there is an integral curve of (3.3) connecting u^L and u^R with the additional property that c is increasing all the way from u^L and u^R (cf. Fig. 3.2). In particular this implies that such a simple c-rarefaction wave never exists when u^L and u^R are located on opposite sides of the transition curve T.

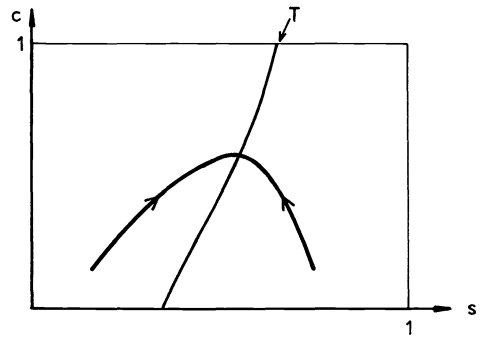


FIG. 3.2. *Rarefaction waves for* $0 < s < 1$.

Finally, consider the case when $s = 0$ or $s = 1$. Then $f_c = 0$ and the eigenvector e^c can be taken to be $(0, 1)$. Hence, the rarefaction curves are given by $s \equiv 0$ or $s \equiv 1$. If $s \equiv 1$ the eigenvalue λ^c is given by

$$\lambda^c(1, c) = \frac{1}{1+h(c)},$$

which is a strictly increasing function of c. Therefore, if $c^L < c^R$ this corresponds to a simple rarefaction wave. Particularly, if $c^T < 1$ (cf. § 2), such a c-rarefaction curve can connect states that are located on opposite sides of the transition curve T. If $s \equiv 0$ the eigenvalue $\lambda^c \equiv 0$. Hence, these "rarefaction curves" correspond to contact discontinuities with speed zero. In order to solve the general Riemann-problem for arbitrary states in $I \times I$ we will allow such discontinuities for arbitrary c^L and c^R in I. We remark however that these solutions are in some sense nonphysical, since states with $s = 0$ and $c > 0$ have no physical interpretation. We will refer to these discontinuities as c-rarefaction waves if $c^L < c^R$ and as c-shock waves if $c^L > c^R$ (cf. § 4).

4. Shock waves and entropy conditions. The purpose of this section is to determine the shock waves of the system (1.1); i.e., for given states $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$ we derive possible discontinuous weak solutions of the system (1.1) of the form

$$u(x, t) = \begin{cases} u^L & \text{if } x/t < \sigma, \\ u^R & \text{if } x/t > \sigma. \end{cases} \tag{4.1}$$

Here σ denotes the shock speed. In order to distinguish the physically meaningful weak solutions of (1.1) we will also require that u satisfies an "entropy condition." This condition will be derived by requiring the shock waves to be evolutionary; i.e., any shock wave must be a limit of traveling wave solutions of associated "viscosity systems."

Any weak solution of (1.1) of the form (4.1) has to satisfy the Rankine-Hugoniot condition given by

$$\begin{aligned} &f(s^R, c^R) - f(s^L, c^L) = \sigma(s^R - s^L), \\ &c^R f(s^R, c^R) - c^L f(s^L, c^L) = \sigma(s^R c^R + a(c^R) - s^L c^L - a(c^L)). \end{aligned} \tag{4.2}$$

Let us first observe that if $c^R = c^L$ then the two equations of (4.2) are identical. Hence, in this case (4.2) reduces to

$$f(s^R, c^L) - f(s^L, c^L) = \sigma(s^R - s^L), \tag{4.3}$$

which is the Rankine-Hugoniot condition for the Buckley-Leverett equation (1.2) with $f(s)=f(s, c^L)$. From the theory for this equation (cf. [11] and § 1) shock waves of the form (4.1), with $c^L=c^R$ and with s^L and s^R satisfying (4.3), are said to satisfy an entropy condition if and only if

$$[f(s)-f(s^L)-\sigma(s-s^L)]\,\text{sign}\,(s^R-s^L)\geqq 0 \tag{4.4}$$

for any s between s^L and s^R. These shock waves will be referred to as *s-shock waves* for the system (1.1). (We remark that the entropy condition (4.4) can be derived, in the same way as below, by requiring the shock wave to be evolutionary.)

The entropy condition (4.4) implies that

$$\lambda^s(u^L)>\sigma>\lambda^s(u^R),$$

and from the definition of λ^c it follows that either

$$\lambda^c(u^L),\ \lambda^c(u^R)\geqq\sigma$$

or

$$\lambda^c(u^L),\ \lambda^c(u^R)\leqq\sigma.$$

Hence, since exactly one characteristic enters the shock on both sides, the s-shock waves satisfy the celebrated Lax entropy condition (cf. [8] and [11]).

In the rest of this section we shall determine the shock waves with $c^L\neq c^R$. These shock waves will be referred to as *c-shock waves.*

By applying the first equation of (4.2), the second equation of (4.2) can be written in the form

$$(c^R-c^L)f(s^L, c^L)=\sigma(c^R-c^L)s^L+\sigma(a(c^R)-a(c^L)). \tag{4.5}$$

By introducing the quantity

$$h_L(c)=\begin{cases}\dfrac{a(c)-a(c^L)}{c-c^L} & \text{if } c\neq c_L,\\ h(c) & \text{if } c=c_L,\end{cases} \tag{4.6}$$

(4.5) can again be written as

$$\sigma=\frac{f(s^L, c^L)}{s^L+h_L(c_R)}.$$

By applying this final equality in the first equation of (4.2) we also obtain that

$$\sigma(s^R+h_L(c^R))=\sigma(s^L+h_L(c^R))+f(s^R, c^R)-f(s^L, c^L)=f(s^R, c^R).$$

Hence, when $c^L\neq c^R$ the Rankine-Hugoniot condition (4.2) can be written in the form

$$\frac{f(s^R, c^R)}{s^R+h_L(c^R)}=\frac{f(s^L, c^L)}{s^L+h_L(c^R)}=\sigma. \tag{4.7}$$

If $s^L=s^R=0$ this condition is satisfied with $\sigma=0$. As we have already seen in § 3 this corresponds to a contact discontinuity with speed zero. In the rest of the discussion we shall therefore assume that $\sigma>0$.

We observe that the value $h_L(c^R)$ is determined from values of c^L and c^R and that, if s^L, c^L and c^R are given, there are at most two values of s^R that satisfy the Rankine-Hugoniot condition (4.7) (cf. Fig. 4.1).

In order to distinguish the physically meaningful weak solutions of (1.1) we shall require that a shock wave be evolutionary. Hence, for any $\varepsilon>0$ we consider the perturbed system

$$\begin{aligned} &s_t+f(s, c)_x=\varepsilon s_{xx},\\ &(sc+a(c))_t+(cf(s, c))_x=\varepsilon(sc+a(c))_{xx}.\end{aligned} \tag{4.8}$$

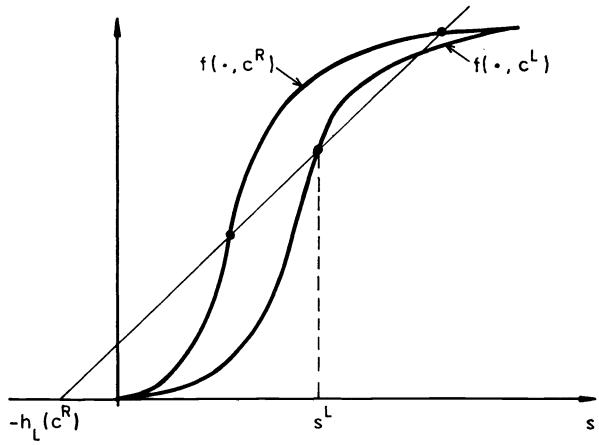


FIG. 4.1. $c^L > c^R$.

The entropy condition consists of requiring the shock wave to be a pointwise limit of traveling wave solutions of (4.8). In the same way as was done in, e.g., [7] or [11, Chap. 24], we obtain that two states $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$, which satisfy the Rankine-Hugoniot condition (4.7) with shock speed σ, correspond to an evolutionary shock wave if and only if the 2×2 system

$$
\begin{aligned}
&\frac{ds}{d\xi} = f(s, c) - \sigma s - (f(s^L, c^L) - \sigma s^L), \\
&\frac{d}{d\xi}(sc + a(c)) = cf(s, c) - \sigma(sc + a(c)) - (c^L f(s^L, c^L) - \sigma(s^L c^L + a(c^L)))
\end{aligned} \tag{4.9}
$$

has a solution $(s(\xi), c(\xi))$ with $(s(-\infty), c(-\infty)) = u^L$ and $(s(+\infty), c(+\infty)) = u^R$. By applying (4.7) and by performing the differentiation in the second equation of (4.9) we rewrite the system in the form

$$
\begin{aligned}
&\frac{ds}{d\xi} = f(s, c) - \sigma(s + h_L(c^R)), \\
&(s + h(c))\frac{dc}{d\xi} = \sigma(c - c^L)(h_L(c^R) - h_L(c)),
\end{aligned} \tag{4.10}
$$

where $h_L(c)$ is defined by (4.6).

We observe that the Rankine-Hugoniot condition (4.7) implies that $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$ are equilibrium points of (4.10). Since $\sigma > 0$ the second equation of (4.10) implies, in particular, that $dc/d\xi < 0$ for any value of c strictly between c^L and c^R. Therefore $c^L > c^R$. For the rest of this discussion we therefore assume that $c^L > c^R$.

Let us first assume that u^R is a state such that

$$\lambda^s(u^R) < \sigma.$$

Then there are at most two possible values of s^L, s^L_- and s^L_+ such that the Rankine-Hugoniot condition (4.7) holds where

$$\lambda^s(s^L_-, c^L) \geqq \sigma \quad \text{and} \quad \lambda^s(s^L_+, c^L) \leqq \sigma$$

(cf. Fig. 4.2).

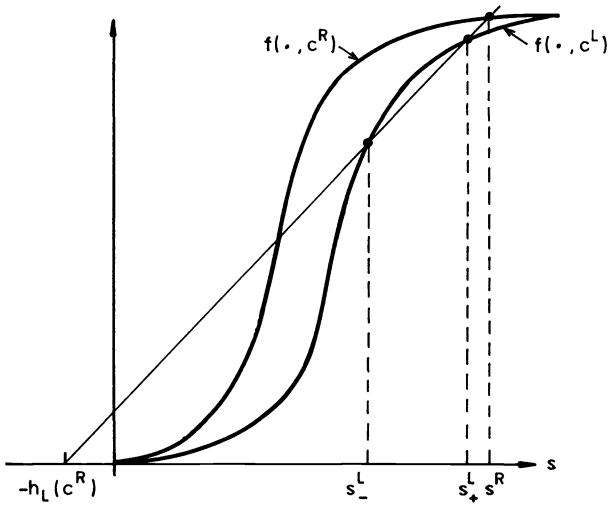


FIG. 4.2

Similarly, for any $c \in [c^R, c^L)$ let $s_-(c)$ and $s_+(c)$ be the two values determined by

$$\frac{f(s_-(c), c)}{s_-(c) + h_L(c^R)} = \sigma = \frac{f(s_+(c), c)}{s_+(c) + h_L(c^R)}$$

and

$$\lambda^s(s_-(c), c) > \sigma > \lambda^s(s_+(c), c).$$

Then $ds/d\xi < 0$ for $0 \leqq s < s_-(c)$ and $s_+(c) < s \leqq 1$, while $ds/d\xi > 0$ for $s_-(c) < s < s_+(c)$ (cf. Fig. 4.3).

It is now easy to see that any trajectory that passes through a point $(s_-(c), c)$, for some $c \in (c^R, c^L)$ has the properties

$$(s(-\infty), c(-\infty)) = (s_-^L, c^L) \quad \text{and} \quad (s(+\infty), c(+\infty)) = (s^R, c^R).$$

Furthermore, any trajectory through $(s_+(c), c)$, for some $c \in (c^R, c^L)$, is such that

$$(s(-\infty), c(-\infty)) = (+\infty, c^L) \quad \text{and} \quad (s(+\infty), c(+\infty)) = (s^R, c^R).$$

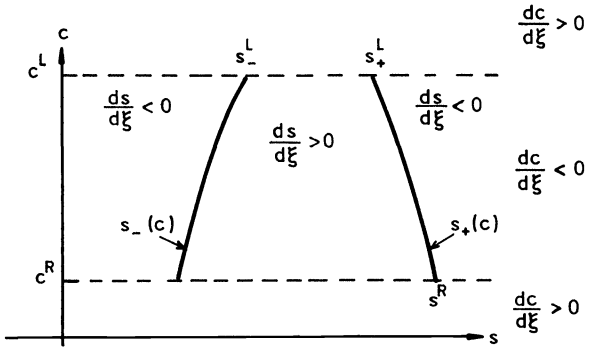


FIG. 4.3

Hence, by continuity, there must be at least one trajectory that separates the two classes and such a trajectory passes from (s_+^L, c^L) to u^R. Therefore we have seen that there exist trajectories from both the points (s_-^L, c^L) and (s_+^L, c^L) to u^R when $\lambda^s(u^R) < \sigma$.

Next consider the case when u^R satisfies

$$\lambda^s(u^R) > \sigma$$

(i.e., $s^R = s_-(c^R)$). By an argument similar to the one above we obtain the existence of a trajectory from (s_-^L, c^L) to u^R. However, if $u^L = (s_+^L, c^L)$ (i.e., $\lambda^s(u^L) < \sigma$) there exists no trajectory from u^L to u^R.

As a conclusion we obtain that two states $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$, with $c^L \neq c^R$, correspond to a c-shock wave of speed σ if they satisfy the Rankine–Hugoniot condition (4.7) and if they satisfy the entropy conditions

$$c^L > c^R \tag{4.11}$$

and

$$\lambda^s(u^R) < \sigma \quad \text{or} \quad \lambda^s(u^L), \lambda^s(u^R) \geqq \sigma. \tag{4.12}$$

Since the function $h(c)$ is strictly decreasing, condition (4.7) implies that (4.11) is equivalent to the eigenvalue/shock speed relation

$$\lambda^c(u^L) > \sigma > \lambda^c(u^R).$$

Therefore, if

$$\lambda^s(u^L), \lambda^c(u^L) > \sigma > \lambda^s(u^R), \lambda^c(u^R),$$

we allow a shock wave where both characteristics on both sides of the shock enter the shock. This shock, which corresponds to $s^L = s_-^L$ and $s^R = s_+(c^R)$, is therefore a shock wave which does not satisfy the Lax entropy condition. We recall, that in the terminology of [10], this is an example of an overcompressive shock. In § 7 below we will discover that this shock cannot be joined by any other wave in a Riemann solution.

5. The Riemann problem. The purpose of the rest of this paper is to construct a unique global solution of the Riemann problem for the system (1.1); i.e., for arbitrary states $u^L, u^R \in I \times I$ we derive a solution of the pure initial value problem for (1.1) with initial condition

$$u(x, 0) = \begin{cases} u^L & \text{if } x < 0, \\ u^R & \text{if } x > 0, \end{cases} \tag{5.1}$$

which consists of a composition of a finite number of simple rarefaction waves and shock waves as described above.

If a left state u^1 can be connected to a right state u^2 by a simple rarefaction wave then the *initial speed* of the wave is $\lambda(u^1)$ and the *final speed* of the wave is $\lambda(u^2)$, where λ is the eigenvalue corresponding to the wave. The initial speed and the final speed of a simple shock wave is defined to be the shock speed σ.

By a *c-wave* we mean a simple c-rarefaction wave or a c-shock wave, while an *s-wave* is any composition of simple s-rarefaction waves and s-shock waves that corresponds to a solution of the Buckley–Leverett equation (1.2) with $f(s) = f(s, c)$ for some $c \in I$. We recall that for any left state $u^1 = (s^1, c)$ and right state $u^2 = (s^2, c)$, where $c \in I$, there is a unique s-wave that connects the two states (cf. the discussion in § 1).

Following [6] we will adopt the notation that $u^1 \xrightarrow{c} u^2$ means that the left state u^1 can be connected to the right state u^2 by a c-wave and we refer to this Riemann solution as $u^1 \xrightarrow{c} u^2$. The notation $u^1 \xrightarrow{s} u^2$ has the analogous meaning for s-waves.

Consider two waves $u^1 \xrightarrow{a} u^2$ and $u^2 \xrightarrow{b} u^3$. Let v_f^a denote the final wave speed of the a-wave and v_i^b the initial wave speed of the b-wave. The two waves are said to be *compatible* if they can be composed to solve the Riemann problem with left state u^1 and right state u^3. Hence the two waves are compatible if and only if

$$v_f^a \leqq v_i^b, \tag{5.2}$$

where we require a strict inequality if both v_f^a and v_i^b are shock speeds. Any solution of the Riemann problem will consist of a sequence of compatible s-waves and c-waves that connects the given left state u^L with the given right state u^R.

The main purpose of this paper is to give a constructive proof of the following existence/uniqueness theorem.

THEOREM 5.1. *For arbitrary states u^L, $u^R \in I \times I$ there exists a unique finite sequence of compatible s-waves and c-waves that generates a solution of the Riemann problem with left state u^L and right state u^R.*

The rest of this paper is devoted to the proof of Theorem 5.1. The following lemma will guarantee that the solution of the Riemann problem is monotone with respect to c; i.e., the solution $u = (s, c)$ of the Riemann problem has the property that the function $c(x, t)$ is a monotone function of x for any $t \geqq 0$.

LEMMA 5.1. *Assume that the three waves*

$$u^L \xrightarrow{c_1} u^1 \xrightarrow{s} u^2 \xrightarrow{c_2} u^R$$

are compatible. Then both the c-waves are rarefaction waves.

Proof. Since the three waves are all compatible we obtain from (5.2) that

$$v_f^1 \leqq v_i^s \leqq v_f^s \leqq v_i^2, \tag{5.3}$$

where v_i^s and v_f^s denote the initial and final speed of the s-wave, respectively, v_f^1 is the final speed of the c_1-wave and v_i^2 is the initial speed of the c_2-wave.

Let $u^1 = (s^1, c)$ and $u^2 = (s^2, c)$ for a suitable $c \in I$ and define

$$\alpha = \frac{f(s^1, c) - f(s^2, c)}{s^1 - s^2}.$$

From the structure of the s-waves (cf. § 1) it follows that

$$v_i^s \leqq \alpha \leqq v_f^s.$$

Therefore, (5.3) implies that

$$v_f^1 \leqq \alpha \leqq v_i^2. \tag{5.4}$$

From §§ 3 and 4 we recall that v_f^1 and v_i^2 are of the form

$$v_f^1 = \frac{f(s^1, c)}{s^1 + h_1} \quad \text{and} \quad v_i^2 = \frac{f(s^2, c)}{s^2 + h_2},$$

where $h_1, h_2 > 0$.

By applying this in (5.4) the inequality can be rewritten in the form

$$h_2 \leqq \frac{f(s^1, c)}{\alpha} - s^1 \leqq h_1. \tag{5.5}$$

However, from (5.5) we immediately obtain the desired result. Assume for example that the c_1-wave is a shock wave and that the c_2-wave is a rarefaction wave. If $s^1 = 0$ (i.e., the c_1-wave is a contact discontinuity), $h_2 = h(c) > 0$ while (5.5) implies that $h_2 \leqq 0$. We can therefore assume that $s^1 > 0$. In this case

$$h_1 = h_L(c) = \frac{a(c) - a(c^L)}{c - c^L} \quad \text{and} \quad h_2 = h(c),$$

which implies that $h_1 < h_2$ since $c < c^L$ and h is strictly decreasing. But this contradicts (5.5). Similar arguments also show that the compositions where the c_1-wave is a rarefaction wave and the c_2-wave is a shock wave or where both c-waves are shock-waves are incompatible. Hence, the only possibility is that both the c-waves are rarefaction waves. In this case $h_1 = h_2 = h(c)$. □

Let $u^L = (s^L, c^L)$ and $u^R = (s^R, c^R)$ denote the left and right state, respectively, of the Riemann problem. The lemma above immediately implies that if $c^L < c^R$, any solution of the Riemann problem will be composed of s-waves and c-rarefaction waves and, if $c^L > c^R$, any solution will be composed of s-waves and a single c-shock wave. Furthermore, if $c^L = c^R$, any solution will consist of a single s-wave. Hence from the theory of the Buckley-Leverett equation (1.2) (cf. § 1) there is a unique solution of the Riemann problem for the system (1.1) when $c^L = c^R$. It is therefore enough to prove Theorem 5.1 when $c^L \neq c^R$.

In § 6 below we treat the case where $c^L < c^R$, while the case $c^L > c^R$ is covered in § 7.

6. The case $c^L < c^R$. Throughout this section we shall consider the Riemann problem for (1.1) with $c^L < c^R$. In this case the purpose is to prove Theorem 5.1.

For any state $u = (s, c) \in \mathscr{R} \cup T$ define the critical value $s^K = s^K(u)$ to be the unique value of s^K such that $(s^K, c) \in \mathscr{L} \cup T$ with the property (cf. Fig. 6.1)

$$\lambda^c(s^K, c) = \lambda^c(u). \tag{6.1}$$

Similarly, if $u = (s, c) \in \mathscr{L}$ let $s^K = s^K(u)$ be the unique s^K such that either $(s^K, c) \in \mathscr{R}$ and that (6.1) holds or, if no such s^K exists, $s^K = \infty$.

As we have already observed above, any solution of the Riemann problem is composed of s-waves and c-rarefaction waves. We first determine the compatible pairs of waves.

LEMMA 6.1. *Assume that $c^L < c^R$.*

(i) *The two waves*

$$u^L \xrightarrow{c} u^M \xrightarrow{s} u^R$$

are compatible if and only if $u^M \in \mathscr{L} \cup T$ and $0 \leqq s^R \leqq s^K(u^M)$.

(ii) *The two waves*

$$u^L \xrightarrow{s} u^M \xrightarrow{c} u^R$$

are compatible if and only if $u^M \in \mathscr{R} \cup T$ and $s^K(u^M) \leqq s^L \leqq 1$.

Proof. Consider part (i) above. The two waves are compatible if and only if

$$\lambda^c(u^M) \leqq v_i^s,$$

where v_i^s is the initial speed of the s-wave. Since $v_i^s \leqq \lambda^s(u^M)$ we therefore derive that $u^M \in \mathscr{L} \cup T$. Furthermore, from the structure of the s-waves it follows that $v_i^s \geqq \lambda^c(u^M)$ if and only if $0 \leqq s^R \leqq s^K(u^M)$.

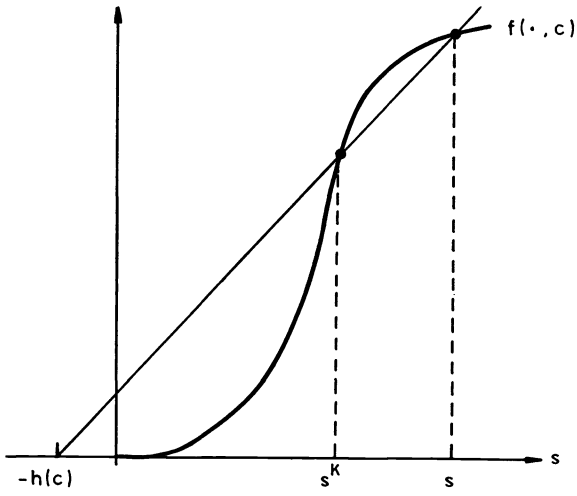


FIG. 6.1

Part (ii) follows by similar arguments. □

The above lemma implies that the three waves

$$u^L \xrightarrow{c_1} u^1 \xrightarrow{s} u^2 \xrightarrow{c_2} u^R \tag{6.2}$$

are compatible if and only if $u^1 \in \mathscr{L}$ and $u^2 \in \mathscr{R}$ with $s^K(u^1) = s^2$ (and $s^K(u^2) = s^1$). Hence, any solution of the Riemann problem will be composed of at most two c-rarefaction waves.

We are now in a position to construct the solution of the Riemann problem for arbitrary states u^L and u^R in $I \times I$. First we consider the case when $u^L \in \mathscr{R} \cup T$.

LEMMA 6.2. *Assume that $c^L < c^R$ and $u^L \in \mathscr{R} \cup T$. Then there exists a unique solution of the Riemann problem consisting of at most four states separated by s-waves and c-rarefaction waves.*

Proof. If $u^R \in \mathscr{R} \cup T$ the solution has the form

$$u^L \xrightarrow{s} u^1 \xrightarrow{c} u^R,$$

where $u^1 = (s^1, c^L) \in \mathscr{R}$ is determined by the c-rarefaction curve which connects u^R to the line $c = c^L$ (cf. Fig. 6.2). (Here and below it might, of course, occur that the s-wave is empty.)

This composition is compatible by part (ii) of Lemma 6.1.

Alternatively, consider the case when $u^R \in \mathscr{L}$. If $s^T(c^R)$ does not exist (cf. § 2), then a compatible composition is given by

$$u^L \xrightarrow{s} (1, c^L) \xrightarrow{c} (1, c^R) \xrightarrow{s} u^R.$$

Otherwise, the composition

$$u^L \to u^T \xrightarrow{s} u^R$$

where $u^T = (s^T(c^R), c^R)$ and $u^L \to u^T$ denotes the solution with left state u^L and right state u^T, is compatible (cf. Fig. 6.3).

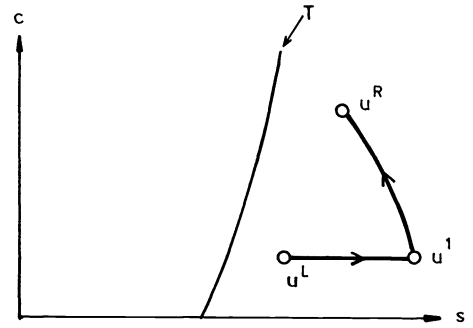


FIG. 6.2

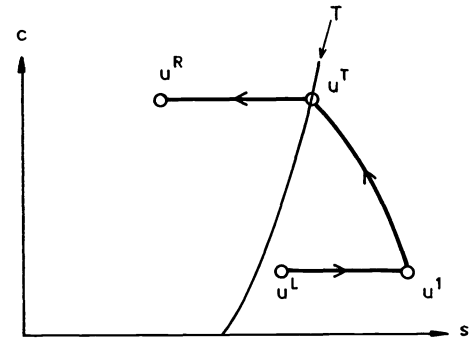


FIG. 6.3

Hence, the existence result of the lemma is established. Verification of the fact that all solutions are unique is straightforward from Lemma 6.1. □

When $u^L \in \mathscr{L}$ the situation is a little more complicated.

LEMMA 6.3. *Assume that $c^L < c^R$ and $u^L \in \mathscr{L}$. Then there exists a unique solution of the Riemann problem consisting of at most five states separated by s-waves and c-rarefaction waves.*

Proof. Let $\Gamma_R \in \mathscr{L} \cup T$ denote the c-rarefaction curve through u^L and let $u^* = (s^*, c^*) \in T$ be the state where Γ_R intersects T. (If no such u^* exists, let $u^* = (s^*, 1)$ be the state where Γ_R intersects the line $c = 1$.) Furthermore, let $\Gamma_K \subset \mathscr{R} \cup T$ be the associated critical curve defined by (cf. Fig. 6.4)

$$\Gamma_K = \{(s^K, c) \in \mathscr{R} \cup T \mid s^K = s^K(s, c) \text{ for } (s, c) \in \Gamma_R\}.$$

From (3.3), defining the c-rarefaction curves, and the definition of $s^K(u)$ it follows that, when $s^K(u) < \infty$,

$$(\lambda^c - \lambda^s)\frac{ds^K}{dc} = f_c + g, \tag{6.3}$$

where $g = g(c) = \lambda^c(s^K, c)(dh/dc)(c)((s^K - s)/(s + h(c)))$, $(s, c) \in \Gamma_R$ and the functions λ^c, λ^s and f_c are all evaluated at (s^K, c). Since $dh/dc < 0$, it follows that $g(c) < 0$ for any $c \in [c^L, c^*)$. By comparing (6.3) with (3.3) we therefore obtain that any c-rarefaction

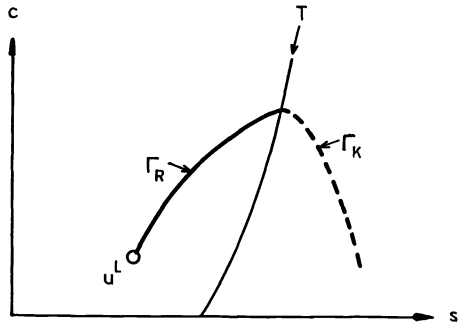


FIG. 6.4. Γ_R *and* Γ_K.

curve in $\mathcal{R}$ can, at most, intersect the critical curve Γ_K once, and that $ds^K/dc < ds/dc$, where $s(c)$ is a c-rarefaction curve (cf. Fig. 6.5).

After this introductory discussion we now construct the solution of the Riemann problem. Assume first that $u^R \in R_1$, where R_1 is the closed region in $I \times I$ bounded by the curves $c = c^L$, $c = c^*$ and Γ_K. In this case Lemma 6.1 implies that the composition

$$u^L \xrightarrow{c} u^1 \xrightarrow{s} u^R,$$

where $u^1 = (s^1, c^R) \in \Gamma_R$ is compatible (cf. Fig. 6.6)

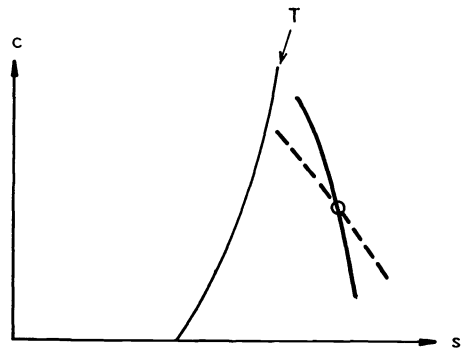


FIG. 6.5. ——— *c-rarefaction curve*; ------ Γ_K.

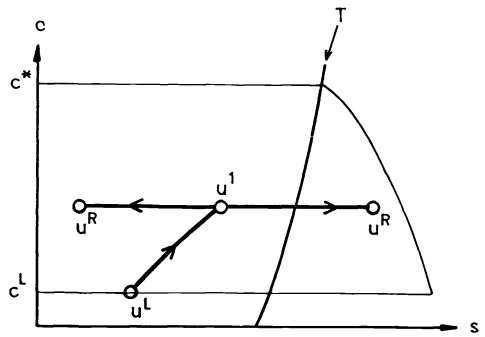


FIG. 6.6

Next assume that $u^R \in R_2$, where R_2 consists of all states in $\mathscr{R} \cup T$ that are not in R_1. Let $u^2 \in \mathscr{R}$ be the state where the c-rarefaction curve through u^R intersects the curve $\Gamma_K \cup \{c = c^L\}$.

If $c^2 = c^L$, Lemma 6.1 implies that the composition

$$u^L \xrightarrow{s} u^2 \xrightarrow{c} u^R$$

is compatible. If $c^2 > c^L$ the composition

$$u^L \xrightarrow{c_1} u^1 \xrightarrow{s} u^2 \xrightarrow{c_2} u^R,$$

where $u^1 = (s^1, c^2) \in \Gamma_R$, corresponds exactly to a compatible composition of the form (6.2). Hence we have constructed a solution for all $u^R \in R_2$ (cf. Fig. 6.7).

Finally, consider the case where

$$u^R \in R_3 = \{u = (s, c) \in \mathscr{L} \mid c > c^*\}.$$

If $s^T(c^R)$ exists the composition

$$u^L \to u^T \xrightarrow{s} u^R,$$

where $u^T = (s^T(c^R), c^R)$ and $u^L \to u^T$ denotes the solution with left state u^L and right state u^T, is compatible (cf. Fig. 6.8).

Otherwise, if $s^T(c^R)$ does not exist, a compatible composition is given by

$$u^L \to (1, c^T) \xrightarrow{c} (1, c^R) \xrightarrow{s} u^R$$

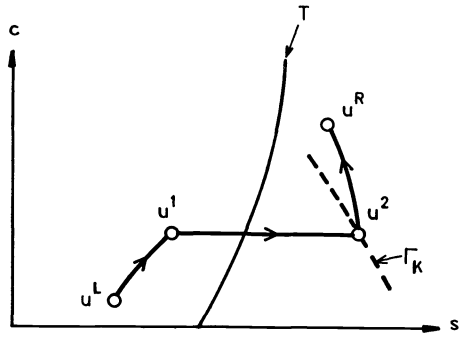


FIG. 6.7

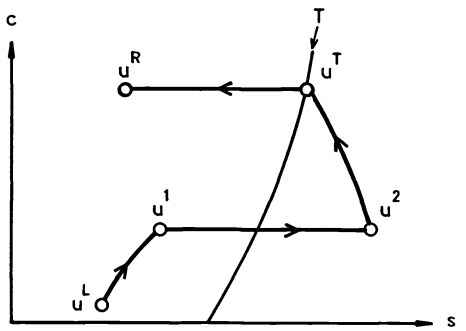


FIG. 6.8

where c^T is the unique value such that $(1, c^T) \in T$ (cf. § 2) and $u^L \to (1, c^T)$ is a solution of a Riemann problem. Hence, the desired existence result is established. Again uniqueness can be verified by applying Lemma 6.1. □

Lemmas 6.2 and 6.3 complete the proof of Theorem 5.1 in the case when $c^L < c^R$.

7. The case $c^L > c^R$. In this section we will complete the proof of Theorem 5.1 by constructing a unique solution of the Riemann problem when $c^L > c^R$. Throughout this section we shall therefore assume that the values of c^L and c^R are *fixed* with $c^L > c^R$.

For any state $u = (s, c) \in I \times I$ we define the "associated shock speed" $\sigma(u)$ by

$$\sigma(u) = \frac{f(s, c)}{s + h_L(c^R)}.$$

Hence, the Rankine-Hugoniot condition (4.7) can be written in the form

$$\sigma(u^L) = \sigma(u^R). \tag{7.1}$$

Similar to the definition given in § 6, for any state $u = (s, c) \in I \times I$, with $\lambda^s(u) \leqq \sigma(u)$, the critical value $s^K = s^K(u)$ is defined to be the unique value of s^K such that $\lambda^s(s^K, c) > \sigma(u)$ and such that (cf. Fig. 7.1)

$$\sigma(s^K, c) = \sigma(u). \tag{7.2}$$

If $\lambda^s(u) > \sigma(u)$, then $s^K(u)$ is either defined to be the unique value s^K such that $\lambda^s(s^K, c) < \sigma(u)$ and such that (7.2) holds or, if no such s^K exists, $s^K = \infty$.

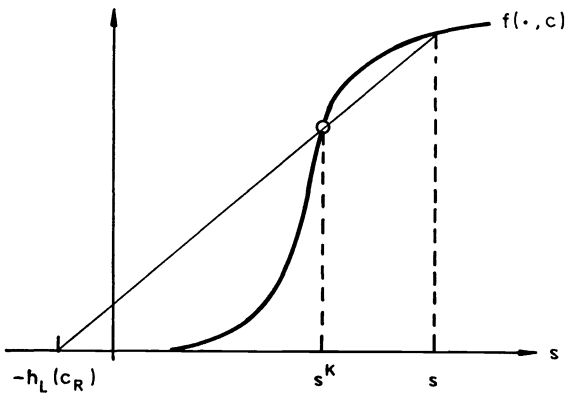


FIG. 7.1

As we have already observed in § 5, any solution of the Riemann problem will be composed of s-waves and one c-shock wave. We first characterize the compatible pairs of waves in the present case.

LEMMA 7.1. *Assume that* $c^L > c^R$.

(i) *The two waves*

$$u^L \xrightarrow{c} u^M \xrightarrow{s} u^R$$

are compatible if and only if $\lambda^s(u^M) \geqq \sigma(u^M)$ *and* $0 \leqq s^R < s^K(u^M)$.

(ii) *The two waves*

$$u^L \xrightarrow{s} u^M \xrightarrow{c} u^R$$

are compatible if and only if $\lambda^s(u^M) \leqq \sigma(u^M)$ *and* $s^K(u^M) < s^R \leqq 1$.

Proof. Consider part (i) above. The two waves are compatible if and only if

$$\sigma \leqq v_s^i, \tag{7.3}$$

where σ is the speed of the c-shock wave, v_s^i is the initial speed of the s-wave and where strict inequality is required if the s-wave starts with a shock wave. But since $\sigma = \sigma(u^M)$ and $v_s^i \leqq \lambda^s(u^M)$ it follows that $\lambda^s(u^M) \geqq \sigma(u^M)$. Furthermore, from the structure of the s-waves it follows that (7.3) holds if and only if $0 \leqq s^R < s^K(u^M)$. Part (ii) follows by similar arguments. □

As a consequence of the lemma above (and Lemma 5.1) we derive in particular that the admissible overcompressive shock discussed in § 4, where

$$\lambda^s(u^L), \lambda^c(u^L) > \sigma > \lambda^s(u^R), \lambda^c(u^R)$$

can never be joined by any other wave in a Riemann solution.

By applying the result of Lemma 7.1 we can now proceed to construct the solution of the Riemann problem.

LEMMA 7.2. *Assume that $c^L > c^R$. Then there exists a unique solution of the Riemann problem consisting of at most four states separated by s-waves and c-shock waves.*

Proof. Consider first the case when $\lambda^s(u^L) \geqq \sigma(u^L)$. Let $u^1 = (s^1, c^R)$ be the unique state such that (cf. Fig. 7.2)

$$\lambda^s(u^1) > \sigma(u^1) = \sigma(u^L).$$

Hence, from the Rankine-Hugoniot condition (7.1) and from the entropy conditions (4.10) and (4.11), the pair (u^L, u^1) corresponds to a c-shock wave with left state u^L, right state u^1 and shock speed $\sigma = \sigma(u^L)$. Furthermore, by part (i) of Lemma 7.1 the composition

$$u^L \xrightarrow{c} u^1 \xrightarrow{s} u^R$$

is compatible if $0 \leqq s^R < s^K(u^1)$. Alternatively, if $s^R > s^K(u^1)$ the composition

$$u^L \xrightarrow{s} u^2 \xrightarrow{c} u^R$$

is compatible, where $u^2 = (s^2, c^L)$ is the state such that

$$\lambda^s(u^2) \leqq \sigma(u^2) = \sigma(u^R).$$

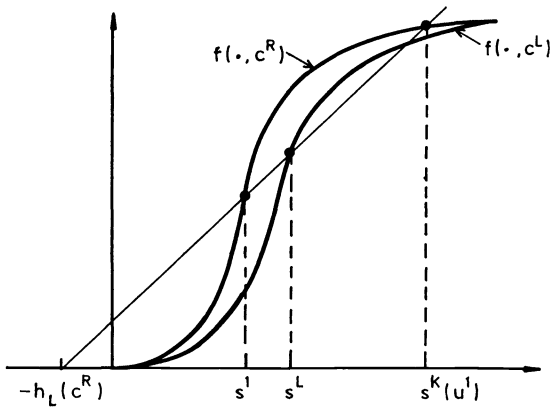


FIG. 7.2

Finally, if $s^R = s^K(u^1)$, the pair (u^L, u^R) corresponds to an admissible overcompressive c-shock (cf. § 4), where $\sigma = \sigma(u^L) = \sigma(u^R)$. Hence, we have constructed the solution of the Riemann problem in all cases when $\lambda^s(u^L) \geqq \sigma(u^L)$ (cf. Fig. 7.3).

Consider next the case when $\lambda^s(u^L) < \sigma(u^L)$ and let $u^* = (s^*, c^L)$ be the unique state such that

$$\sigma(u^*) = \lambda^s(u^*).$$

Furthermore, let $u^1 = (s^1, c^R)$ be determined by

$$\lambda^s(u^1) > \sigma(u^1) = \sigma(u^*)$$

and let $s^K = s^K(u^1)$ (cf. Fig. 7.4).

If $s^R \geqq s^K$ a solution of the Riemann problem is given by the composition

$$u^L \xrightarrow{s} u^2 \xrightarrow{c} u^R$$

where $u^2 = (s^2, c^L)$ is the unique state such that $\lambda^s(u^2) \leqq \sigma(u^2)$ and such that the pair (u^2, u^R) corresponds to a c-shock wave (cf. Fig. 7.5).

If $s^R < s^K$ a solution is given by the composition

$$u^L \xrightarrow{s} u^* \xrightarrow{c} u^1 \to u^R,$$

which is compatible by Lemma 7.1 (cf. Fig. 7.6).

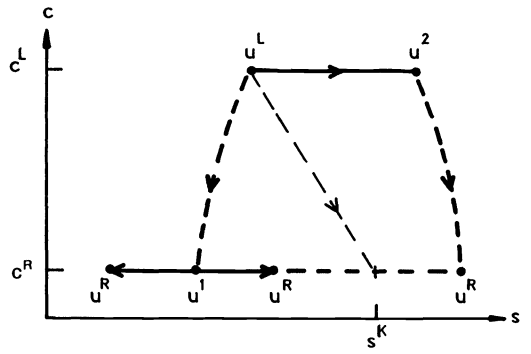


FIG. 7.3

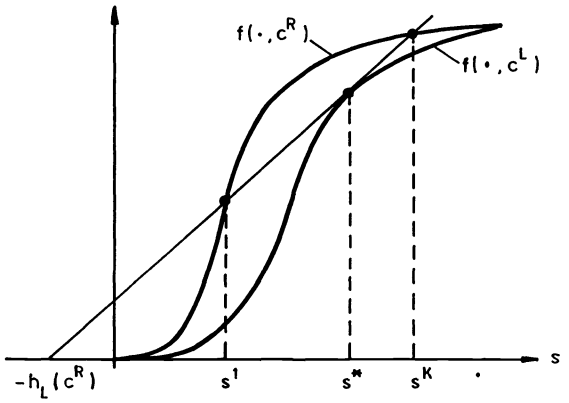


FIG. 7.4

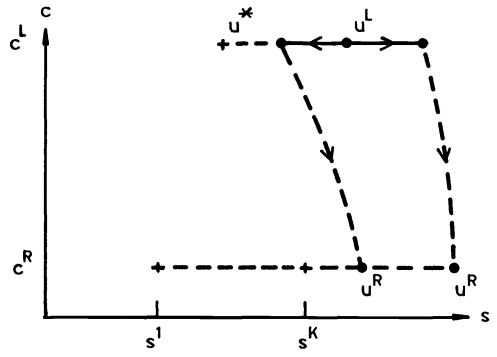


FIG. 7.5

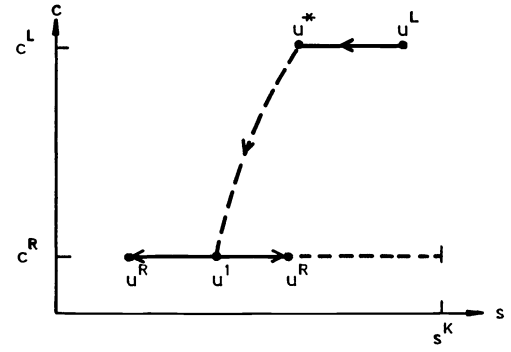


FIG. 7.6

Hence the desired existence result is established, and uniqueness can easily be verified by applying Lemma 7.1. □

As a consequence of the three Lemmas 6.2, 6.3, and 7.2 the proof of Theorem 5.1 is now completed. We observe that the existence proof is constructive; i.e., we have constructed an algorithm for the solution of the global Riemann problem for the system (1.1).

8. Numerical experiments. Based on the constructive proof of Theorem 5.1 a computer program is developed that solves the global Riemann problem for the model (1.1). In order to solve nonlinear equations obtained from the Rankine-Hugoniot condition (4.7) and the ordinary differential equations involved in the determination of the rarefaction waves, standard numerical methods are used.

The purpose of the numerical experiments presented below is to illustrate some typical behavior in the exact solution of the Riemann problem and to show that sometimes this behavior is not easily detected by means of finite difference schemes. In the examples presented below we have used the algebraic expressions

$$f(s, c) = \frac{s^2}{s^2 + (0.5 + 100c)(1-s)^2} \tag{8.1}$$

and

$$a(c) = \frac{0.2c}{1+100c}. \tag{8.2}$$

The function f is graphed in Fig. 8.1 for different values of c uniformly distributed between 0.00 and 0.01.

Example 8.1. In this example we calculate the solution of the Riemann problem with $u^L = (s^L, c^L) = (1.0, 0.01)$ and $u^R = (s^R, c^R) = (0.0, 0.0)$. This models a situation where the porous medium is initially 100 percent saturated with oil. Then water containing polymer is injected in order to displace the oil.

The structure of the solution of the Riemann problem in this case is illustrated by Fig. 7.6. The exact solution is shown on Fig. 8.2.

We observe that a bank of polymer-free water separates the polymer from the region with 100 percent oil. We remark that this bank forms as a consequence of the adsorption term $a(c)$ in the system (1.1).

One of the purposes of polymer injection is to increase the viscosity of the aqueous phase in order to reduce viscous fingering. The bank of pure water observed above might therefore decrease the desired efficiency of the displacement process.

Example 8.2. Throughout this example we consider the Riemann problem with $u^L = (s^L, c^L) = (0.9, 0.007)$ and $c^R = 0.003$ fixed and where s^R is close to the critical value s^K. These cases are illustrated in Fig. 7.5 ($s^R \geqq s^K$) and Fig. 7.6 ($s^R < s^K$). Examples of the s-component of the solution is shown in Fig. 8.3 and Fig. 8.4 below. Since the values s^* and s^1 are independent of s^R for $s^R < s^K$, this shows that, if the solution of the Riemann problem is considered pointwise, it is discontinuous with respect to s^R close to s^K. However, since the width of the constant solution s^1 is

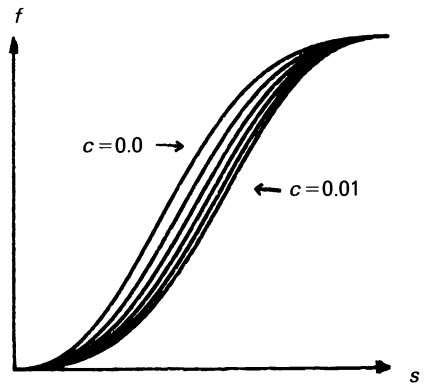


FIG. 8.1

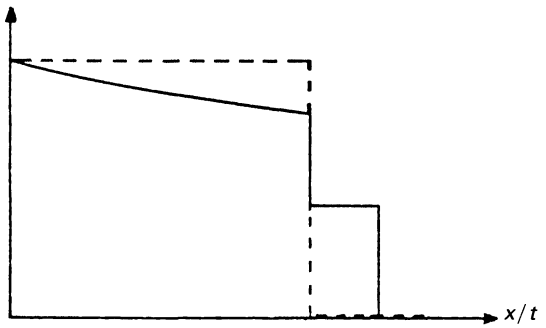
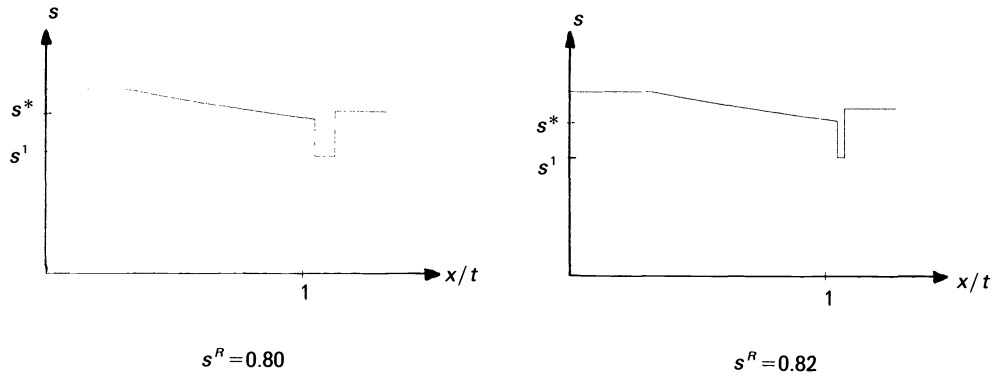


FIG. 8.2. ——— s; ------ $c \times 100$.



$s^R = 0.80$ $s^R = 0.82$

FIG. 8.3. $s^R < s^K$.

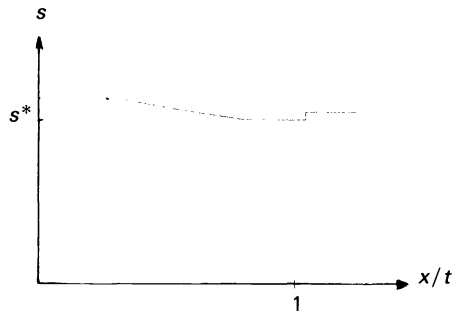


FIG. 8.4. $s^R = 0.84 > s^K$.

vanishing as s^R tends to s^K from below, the solution is continuous with respect to s^R in L^1 norm.

Example 8.3. Consider next the Riemann problem with $u^L = (s^L, c^L) = (0.45, 0.0)$ and $u^R = (s^R, c^R) = (0.20, 0.01)$. This corresponds to the case illustrated in Fig. 6.8 where the maximum number of intermediate states is present in the Riemann solution. The exact solution is shown in Fig. 8.5.

In order to test the efficiency of a standard difference method we have calculated the solution of this Riemann problem with the standard 1 order upwind scheme using a uniform grid. We have used $\Delta t/\Delta x = 12/25$ which satisfies the CFL stability condition.

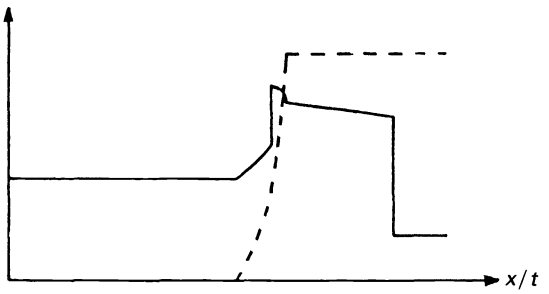


FIG. 8.5. ——— s; ------ $c \times 100$.

An approximation of the solution at $t = 1$ is calculated with $\Delta t = 0.04$ and the approximate saturation is compared with the exact saturation in Fig. 8.6.

As we observe the numerical method seems to neglect certain effects reflected in the exact solution. In order to investigate these phenomena further, similar numerical solutions, with decreasing time steps, have been calculated. In Fig. 8.7 we have plotted the different numerical paths in (s, c)-space together with the exact path. We have used $\Delta t = 0.04$, 0.016, 0.008, 0.004, 0.0016.

The results seem indeed to indicate that the solutions obtained by the upwind scheme converge to the exact solution. However, it might be difficult to give an accurate physical interpretation of the process from the numerical results.

A comparison between the numerical solution of s with $\Delta t = 0.0016$ and the exact solution (cf. Fig. 8.4) is shown in Fig. 8.8.

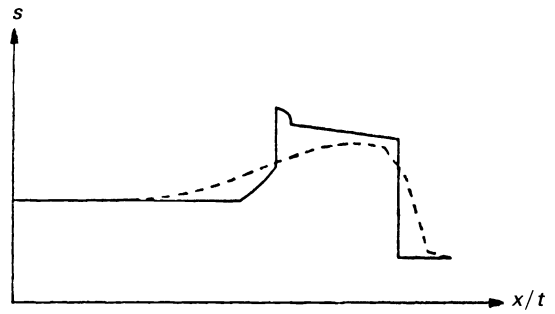


FIG. 8.6. ——— *exact solution of s*; ------ *numerical solution of s.* $\Delta t = 0.04$.

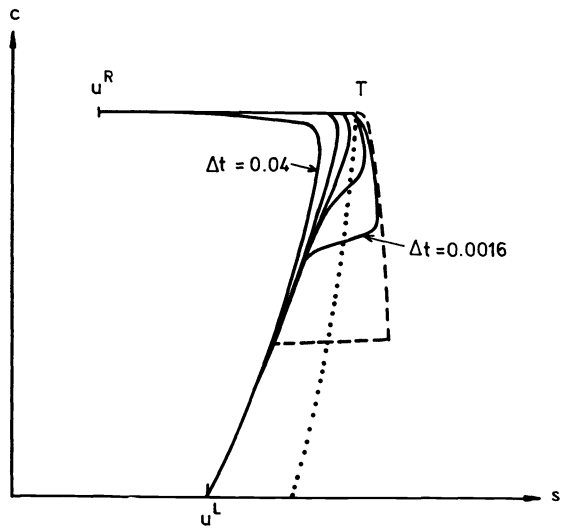


FIG. 8.7. ——— *numerical paths*; ------ *exact path*; ····· *T.*

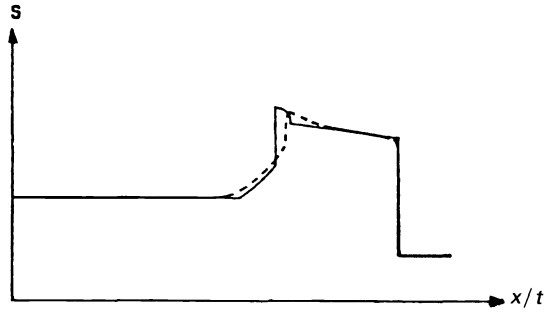


FIG. 8.8. ——— *exact solution of s*; ------ *numerical solution of s.* $\Delta t = 0.0016$.

Acknowledgment. The authors would like to thank Aslak Tveito for carrying out the computations presented in § 8.

REFERENCES

[1] A. I. CHORIN, *Random choice solutions of hyperbolic systems*, J. Comput. Phys., 22 (1976), pp. 517–533.
[2] P. CONCUS AND W. PROSKUROWSKI, *Numerical solutions of a nonlinear hyperbolic equation by the random choice method*, J. Comput. Phys., 30 (1979), pp. 153–166.
[3] I. GELFAND, *Some Problems in the Theory of Quasilinear Equations*, American Mathematical Society Translation Ser. 2, 1963, pp. 295–381.
[4] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.
[5] S. K. GODUNOV, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Math. USSR Sb., 47 (1959), pp. 271–290.
[6] E. ISAACSON, *Global solution of a Riemann problem for a non-strictly hyperbolic system of conservation laws arising in enhanced oil recovery*, Rockefeller University, New York, NY, preprint.
[7] B. KEYFITZ AND H. KRANZER, *A system of non-strictly hyperbolic conservation laws arising in elasticity theory*, Arch. Rational Mech. Anal., 72 (1980), pp. 219–241.
[8] P. D. LAX, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[9] G. A. POPE, *The application of fractional flow theory to enhanced oil recovery*, Soc. Pet. Engrg. J., 20 (1980), pp. 191–205.
[10] D. G. SCHAEFFER AND M. SHEARER, *Riemann problems for nonstrictly hyperbolic* 2×2 *systems of conservation laws*, Trans. Amer. Math. Soc., to appear.
[11] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1982.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

005

DISCONTINUOUS SOLUTIONS OF THE LINEARIZED, STEADY STATE, COMPRESSIBLE, VISCOUS, NAVIER–STOKES EQUATIONS*

R. BRUCE KELLOGG†

Abstract. The compressible steady state viscous Navier–Stokes equations in two space dimensions are considered. The equations are linearized around a given ambient flow field. It is shown that the linearized equations are not, in general, elliptic. Jump conditions across a possible curve of discontinuities of a solution of the linearized system are derived. In a particular case, a discontinuous solution of the linearized system is constructed.

Key words. compressible flow, Navier–Stokes equations, discontinuous solutions

AMS(MOS) subject classifications. 76N10, 35Q10

1. Introduction. A complete mathematical understanding of the equations of fluid dynamics has not yet been achieved. Some insight concerning these equations may be obtained from the linearized equations. In this paper we consider the system of equations obtained by linearizing the two-dimensional, compressible steady state viscous Navier–Stokes equations around a smooth, nonzero ambient flow field. This linearized system is not, in general, elliptic in the sense of Agmon, Douglis, and Nirenberg (ADN). (In the case of incompressible flow, the linearized equations are ADN elliptic.) It is therefore possible that the linearized equation have solutions with interior discontinuities. The main goal of the paper is to exhibit a discontinuous solution of these equations in the particular case when the ambient flow field is a constant nonzero flow. The discontinuous solution is constructed with the help of the Fourier transform. In addition, we derive jump conditions that a weak solution of the equations must satisfy if the weak solution has a curve Γ of discontinuities. The jump conditions permit a jump in the normal stress across Γ which is compensated by a jump in the pressure across Γ. Also, the jump conditions require that the temperature and its first derivatives be continuous across Γ. Finally, we consider the uniqueness of the discontinuous solution. The existence of discontinuous solutions may raise the issue of whether some further physical principle is required to specify the solution, analogous to the entropy condition in the nonlinear inviscid case. We investigate this in the particular case when the ambient flow field is constant. We show that an appropriate set of boundary conditions determines the solution uniquely. The boundary conditions, which are also used in [3], require the specification of velocity and temperature on the boundary of the region, and the specification of pressure on that part of the boundary for which the ambient flow enters the region.

The interior discontinuity in the solution that we construct arises from a jump in the specified pressure on the boundary of the region. The jump in the boundary pressure propagates to a jump in the pressure across the streamline of the ambient flow field that emanates from the boundary discontinuity. The strength of the discontinuity increases with smaller viscosity, but decays as one moves along the streamline away from the boundary. The rate of decay of the strength of the discontinuity is given by

* Received by the editors August 29, 1985; accepted for publication (in revised form) May 21, 1987.

† Applied Mathematics Branch (R44), Naval Surface Weapons Center, Silver Springs, Maryland 20910 and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The work of this author was supported in part by the Naval Surface Weapons Center Independent Research Funds and in part by National Science Foundation grant INT-8517582.

the dimensionless quantity $x\rho_0/\mu U_0(\partial\rho/\partial P)$, where x is the distance along the ambient streamline, μ is the coefficient of viscosity, ρ_0 and U_0 are the (constant) density and velocity of the ambient flow, and the derivative of density with respect to pressure is evaluated at constant temperature. In the incompressible case, $\partial\rho/\partial P = 0$, and the rate of decay of the strength of the discontinuity is infinite.

Our work is related to a recent paper of Geymonat and Leyland [3], in which the time-dependent linearized compressible Navier-Stokes equations are discussed. The paper [3] provides an appropriate set of boundary conditions for the problem, gives the existence of a solution in a functional analytic setting, and has suggested some of the arguments used here. Our results may also be contrasted with those in a forthcoming paper of Hoff and Liu [4]. Reference [4] treats the full nonlinear time-dependent equations in one space variable. It is shown that if there is a discontinuity in the initial conditions, the discontinuity persists for all time, but the strength of the discontinuity decays exponentially with time. In the result of [4], the boundary conditions are continuous. Thus, these results are not inconsistent with our finding that, in the linear case, a discontinuity in the boundary conditions produces a discontinuity in the solution of the steady state problem.

In § 2 we derive the linearized equations, discuss the lack of ellipticity of the system, and derive the jump conditions across a putative curve of discontinuity. In § 3 we consider the case of uniform ambient flow. Using an energy argument, we derive an existence and uniqueness theorem for an appropriate boundary value problem associated with the linearized equations. Finally, we give the construction of a discontinuous solution. In § 4 we give some conclusions.

2. The linearized equations. We first write the two-dimensional compressible Navier-Stokes equations. We take the primary unknowns to be the velocity vector $(U(x, y), V(x, y))$, the pressure $P(x, y)$ and the temperature $\Theta(x, y)$, and we let

$$\rho(P, \Theta) = \text{density},$$

$$\varepsilon(P, \Theta) = \text{internal energy},$$

$$Q = \tfrac{1}{2}(U^2 + V^2), \qquad e = \rho\varepsilon + \rho Q,$$

$$\sigma_{11} = 2\mu U_x + \lambda(U_x + V_y),$$

$$\sigma_{12} = \sigma_{21} = \mu(U_y + V_x),$$

$$\sigma_{22} = 2\mu V_y + \lambda(U_x + V_y).$$

The viscosity coefficients λ, μ, and the thermal conductivity, κ, are assumed to be constant, and the Latin subscripts denote partial derivatives. With this notation, the compressible flow equations may be written, in divergence form, as

$$-\sigma_{11,x} - \sigma_{12,y} + (\rho U^2)_x + (\rho UV)_y + P_x = 0,$$

$$-\sigma_{12,x} - \sigma_{22,y} + (\rho UV)_x + (\rho V^2)_y + P_y = 0,$$

$$(\rho U)_x + (\rho V)_y = 0,$$

$$-\kappa\Delta\Theta - (U\sigma_{11})_x - (V\sigma_{12})_x - (U\sigma_{12})_y - (V\sigma_{22})_y,$$

$$+(Ue + UP)_x + (Ve + VP)_y = 0.$$

In the usual manner, after substituting the continuity equation into the momentum equations, and the momentum equations into the energy equation, we obtain

$$(2.1a) \qquad -\sigma_{11,x}-\sigma_{12,y}+\rho UU_x+\rho VU_y+P_x=0,$$

$$(2.1b) \qquad -\sigma_{12,x}-\sigma_{22,y}+\rho UV_x+\rho VV_y+P_y=0,$$

$$(2.1c) \qquad (\rho U)_x+(\rho V)_y=0,$$

$$(2.1d) \qquad -\kappa\Delta\Theta-U_x\sigma_{11}-(V_x+U_y)\sigma_{12}-V_y\sigma_{22}+\rho(U\varepsilon_x+V\varepsilon_y)+PU_x+PV_y=0.$$

We take the system (2.1) as our starting point. Suppose $U(x, y)$, $V(x, y)$, $P(x, y)$, $\Theta(x, y)$ are a solution to (2.1). We shall study the linearized system, linearizing around the "ambient" flow field U, V, P, Θ. Denote the linearized variables by u, v, p, θ, let $q=[u, v, p, \theta]^T$ be the vector of unknowns, and set $\rho_1=\partial\rho(P, \Theta)/\partial P$, $\rho_2=\partial\rho(P, \Theta)/\partial\Theta$, with similar notation for ε_1 and ε_2. The linearized equations are

$$(2.2a) \qquad L_1q\equiv-(2\mu+\lambda)u_{xx}-\mu u_{yy}-(\mu+\lambda)v_{xy}+\rho Uu_x+\rho Vu_y+p_x+A=0,$$

$$(2.2b) \qquad L_2q\equiv-\mu v_{xx}-(2\mu+\lambda)v_{yy}-(\mu+\lambda)u_{xy}+\rho Uv_x+\rho Vv_y+p_y+B=0,$$

$$(2.2c) \qquad L_3q\equiv\rho u_x+\rho v_y+\rho_1(Up_x+Vp_y)+\rho_2(U\theta_x+V\theta_y)+C=0,$$

$$(2.2d) \qquad \begin{aligned} L_4q\equiv&-\kappa\Delta\theta-\sigma_{11}u_x-\sigma_{12}(v_x+u_y)-\sigma_{22}v_y\\ &-U_x[2\mu u_x+\lambda(u_x+v_y)]-\mu(U_y+V_x)(u_y+v_x)\\ &-V_y[2\mu v_y+\lambda(u_x+v_y)]\\ &+\rho\varepsilon_1(Up_x+Vp_y)+\rho\varepsilon_2(U\theta_x+V\theta_y)+P(u_x+v_y)+D=0. \end{aligned}$$

In these equations, A, B, C, and D contain only undifferentiated terms in the linearized variables. Each term in A, B, C, or D contains derivatives of the ambient variables. Thus, in the particular case that the ambient field has uniform flow, these lowest order terms vanish. This fact will be used in the following section.

We first consider the ellipticity of the system (2.2).

LEMMA 1. *The system defined by the operators L_j in* (2.2) *is not, in general, elliptic in the sense of ADN.*

Proof. We use the Volevich criterion, as described in [1]. For this, we replace the derivatives in (2.2) by symbolic multipliers, ξ and η and their powers, and we consider the determinant

$$G=\begin{bmatrix} -(2\mu+\lambda)\xi^2-\mu\eta^2+\alpha_1^1 & -(\mu+\lambda)\xi\eta+\alpha_2^1 & \xi+\alpha_3^0 & \alpha_4^0 \\ -(\mu+\lambda)\xi\eta+\beta_1^1 & -\mu\xi^2-(2\mu+\lambda)\eta^2+\beta_2^1 & \eta+\beta_3^0 & \beta_4^0 \\ \rho\xi+\gamma_1^0 & \rho\eta+\gamma_2^0 & \rho_1(U\xi+V\eta)+\gamma_3^0 & \gamma_4^1 \\ \delta_1^1 & \delta_2^1 & \delta_3^1 & -\kappa(\xi^2+\eta^2)+\delta_4^1 \end{bmatrix}.$$

The α, β, γ, δ are polynomials in ξ and η, which may be determined from (2.2), but whose exact form is not important for the proof. The maximum degree in each of these polynomials is given by the superscript; thus, α_1^1 is a linear polynomial in ξ and η, α_3^0 is independent of ξ and η, etc. When the determinant G is written out as a sum of terms, each term is a polynomial in ξ and η. The degree of each term is $\leqq 7$, since each term in row 3 has degree $\leqq 1$. If $\rho_1\neq 0$, the product of the diagonal entries has degree 7. Hence the leading terms in G cannot be nonzero for not-zero (ξ, η), so the Volevich criterion is not satisfied and the system is not elliptic.

We remark that if the expansion of G is examined further, the terms of order 7 are found to be

$$-\rho_1\kappa\mu(2\mu+\lambda)(\xi^2+\eta^2)^3(U\xi+V\eta).$$

Thus, the streamline of the ambient flow field seems to play the role of a characteristic direction for this system. On the other hand, if $\rho_1=0$, we must examine terms of order 6 to determine the ellipticity of the system.

Since the system (2.2) is not elliptic, it is of interest to ask whether a solution of this system can be discontinuous. For this, we must define the notion of a weak solution of (2.2). Let $\Omega \subset R^2$ be an open set. We say that $q=[u, v, p, \theta]$ is a weak solution of (2.2) in Ω if u, v, p, and θ are integrable, and if for any $\phi \in C_0^\infty(\Omega)$, we have

$$\iint_\Omega \{u[-(2\mu+\lambda)\phi_{xx}-\mu\phi_{yy}-(\rho U\phi)_x-(\rho V\phi)_y]-(\mu+\lambda)v\phi_{xy}-p\phi_x$$

$$+A\phi\}\,dx\,dy=0, \tag{2.3a}$$

$$\iint_\Omega \{v[-\mu\phi_{xx}-(2\mu+\lambda)\phi_{yy}-(\rho U\phi)_x-(\rho V\phi)_y]-(\mu+\lambda)u\phi_{xy}-p\phi_y$$

$$+B\phi\}\,dx\,dy=0, \tag{2.3b}$$

$$\iint_\Omega \{-(\rho\phi)_x u-(\rho\phi)_y v-p[(\rho_1 U\phi)_x+(\rho_1 V\phi)_y]-\theta[(\rho_2 U\phi)_x+(\rho_2 V\phi)_y]$$

$$+C\phi\}\,dx\,dy=0, \tag{2.3c}$$

$$\begin{aligned}\iint_\Omega \{&\theta[-\kappa\Delta\phi-(\rho\varepsilon_2 U)_x-(\rho\varepsilon_2 V)_y]+u[(\sigma_{11}\phi)_x+(\sigma_{12}\phi)_y\\ &+2\mu(U_x\phi)_x+\lambda(U_x\phi)_x+\mu((U_y+V_x)\phi)_y+\lambda(V_y\phi)_x-(P\phi)_x]\\ &+v[(\sigma_{12}\phi)_x+(\sigma_{22}\phi)_y+\lambda(U_x\phi)_y+\mu((U_y+V_x)\phi)_x\\ &+2\mu(V_y\phi)_y+\lambda(V_y\phi)_y-(P\phi)_y]\\ &-p[(\rho\varepsilon_1 U\phi)_x+(\rho\varepsilon_1 V\phi)_y]+D\phi\}\,dx\,dy=0.\end{aligned} \tag{2.3d}$$

Suppose we are given a smooth curve $\Gamma=(x(s), y(s))$ in Ω which divides Ω into 2 pieces, Ω_1 and Ω_2. Suppose that the parameter s is arc length, so that $\underline{n}(s)=[\dot{y}(s), -\dot{x}(s)]^T$ is the unit normal vector along Γ. Suppose $\underline{n}(s)$ points into the region Ω_2, so $\underline{n}(s)$ is an outward pointing normal from the region Ω_1. Let $q=[u, v, p, \theta]$ be a weak solution of (2.2) such that the functions u, v, p, θ are smooth in each of the regions Ω_1, Ω_2 but u, v, p, θ and their derivatives may have jumps across Γ. Let u_1, u_2, u_{x1}, u_{x2}, etc. denote the one-sided limits of u, u_x, etc., on Γ, and let $\delta u=u_2-u_1$, $\delta u_x=u_{x2}-u_{x1}$ denote the jumps in these quantities. Also, let $u_t=\dot{x}u_x+\dot{y}u_y$, $u_n=\dot{y}u_x-\dot{x}u_y$, denote the normal and tangential derivatives of u on Γ, with a similar meaning for v_t, v_n, etc. With these assumptions and notation, we can state the jump conditions for our solution on Γ. The proof is a standard integration by parts argument, as is found, for example, in [2, V.1.3].

THEOREM 1. *Let the domain Ω be separated into subdomains Ω_1, Ω_2, by a smooth curve Γ. Let $q=[u, v, p, \theta]$ be a weak solution of* (2.1). *Suppose q is smooth in each subdomain Ω_i, with smooth one-sided limits on Γ. Suppose q is discontinuous on Γ. Then Γ is a steamline of the ambient flow field, and q satisfies on Γ the jump conditions*

$$\delta u=\delta v=\delta u_t=\delta v_t=0, \tag{2.4a}$$

$$\delta\theta=\delta\theta_x=\delta\theta_y=0, \tag{2.4b}$$

$$[\mu+(\mu+\lambda)\dot{y}^2]\delta u_n-\dot{x}\dot{y}(\mu+\lambda)\delta v_n=\dot{y}\delta p, \tag{2.4c}$$

$$[\mu+(\mu+\lambda)\dot{x}^2]\delta v_n-\dot{x}\dot{y}(\mu+\lambda)\delta u_n=-\dot{x}\delta p. \tag{2.4d}$$

Proof. Pick ϕ to have support in Ω_1. Then since u, v and p are smooth in Ω_1, (2.3a) may be integrated by parts twice to obtain $\iint \phi L_1 q\,dx\,dy=0$. Since this holds

for all such ϕ, $L_1 q = 0$ in Ω_1. A similar argument applies to Ω_2, and to the other equations of (2.2). Next, pick ϕ to have support including a portion of Γ. Writing the integral in (2.3a) as the sum of integrals over Ω_1 and Ω_2, and integrating each of these by parts, we obtain

$$\int_\Gamma \{-(2\mu+\lambda)\dot{y}[\phi_x \delta u - \phi \delta u_x] + \mu \dot{x}[\phi_y \delta u - \phi \delta u_y] + (\mu+\lambda)[\dot{x}\phi_x \delta v + \dot{y}\phi \delta v_y] - \rho\phi\delta u(U\dot{y} - V\dot{x}) - \dot{y}\phi\delta p\}\, ds = 0. \tag{2.5a}$$

We use the relations

$$\phi_x = \dot{x}\phi_t + \dot{y}\phi_n, \qquad \phi_y = \dot{y}\phi_t - \dot{x}\phi_n \tag{2.6}$$

in (2.5a). Since ϕ_n may be chosen independent of ϕ and ϕ_t on Γ, and since (2.5a) must hold for all test functions ϕ, we may set the coefficient of ϕ_n equal to zero. This gives, after some algebra,

$$[(2\mu+\lambda)\dot{y}^2 + \mu\dot{x}^2]\delta u - (\mu+\lambda)\dot{x}\dot{y}\delta v = 0. \tag{2.7a}$$

Writing the integral in (2.3b) as a sum of integrals over Ω_1 and Ω_2, and integrating each of these by parts, we obtain

$$\int_\Gamma \{-\mu\dot{y}[\phi_x \delta v - \phi \delta v_x] + (2\mu+\lambda)\dot{x}[\phi_y \delta v - \phi \delta v_y] + (\mu+\lambda)[\dot{x}\phi_x \delta u + \dot{y}\phi \delta u_y] - \rho\phi\delta v(U\dot{y} - V\dot{x}) + \dot{x}\phi\delta p\}\, ds = 0. \tag{2.5b}$$

Using (2.6) in (2.5b) we derive, in a manner similar to the derivation of (2.7a),

$$-\dot{x}\dot{y}(\mu+\lambda)\delta u + [\mu\dot{y}^2 + (2\mu+\lambda)\dot{x}^2]\delta v = 0. \tag{2.7b}$$

The system (2.7a), (2.7b) forms a system of two homogeneous linear equations in the two unknowns δu, δv. The determinant of this system is $\mu(2\mu+\lambda) \neq 0$. Hence $\delta u = \delta v = 0$ on Γ. From this it follows that $\delta u_t = \delta v_t = 0$, so (2.4a) has been verified. Using (2.4a) in (2.5a), we obtain

$$\int_\Gamma \phi\{(2\mu+\lambda)\dot{y}\delta u_x - \mu\dot{x}\delta u_y + (\mu+\lambda)\dot{y}\delta v_y - \dot{y}\delta p\}\, ds = 0.$$

Since this holds for all test functions ϕ, the quantity in braces must vanish. Using (2.4a) again, we find that $\delta u_x = \dot{y}\delta u_n$, $\delta u_y = -\dot{x}\delta u_n$, $\delta v_y = -\dot{x}\delta v_n$, and we conclude that (2.4c) holds. In a similar manner, (2.4d) is derived from (2.5b). We next consider the jump conditions coming from the energy equation. Writing the integral in (2.3d) as a sum of integrals over Ω_1 and Ω_2, integrating each of these by parts, and using the fact that $\delta u = \delta v = 0$, we obtain

$$\int_\Gamma \{-\kappa[\phi_n \delta\theta - \phi\delta\theta_n] - \rho\varepsilon_2[\dot{y}U\delta\theta - \dot{x}V\delta\theta] - \rho\varepsilon_1\phi\delta p[U\dot{y} - V\dot{x}]\}\, ds = 0.$$

Since ϕ_n may be chosen independently of ϕ, we obtain $\delta\theta = 0$ and

$$\kappa\delta\theta_n - \rho\varepsilon_1[U\dot{y} - V\dot{x}]\delta p = 0. \tag{2.8}$$

Finally, we consider the jump condition coming from the continuity equation. Writing the integral in (2.3c) as a sum of integrals over Ω_1 and Ω_2, integrating each of these

by parts, and using the fact that $\delta u = \delta v = \delta\theta = 0$, we obtain

$$\int_\Gamma \rho_1 \phi [U\dot{y} - V\dot{x}]\delta p \, ds = 0,$$

so we get

$$\rho_1[U\dot{y} - V\dot{x}]\delta p = 0. \tag{2.9}$$

If $U\dot{y} - V\dot{x} \neq 0$, then $\delta p = 0$. So, from (2.8), $\delta\theta_n = 0$, and from (2.4c), (2.4d), $\delta u_n = \delta v_n = 0$. In this case, there is no jump. If $U\dot{y} - V\dot{x} = 0$ so that Γ lies on a streamline of the ambient flow, from (2.8) we get $\delta\theta_n = 0$. Since $\delta\theta_t = 0$, we have $\delta\theta_x = \delta\theta_y = 0$. This establishes (2.4b) and completes the proof of the theorem.

3. Discontinuous solutions. We have seen that a generalized solution of (2.2) must satisfy (2.4) across a curve $(x(s), y(s))$ of discontinuity. The jump conditions (2.4) are analogous to the Rankine-Hugoniot condition for the discontinuous solution of a nonlinear conservation law. It may be asked if (2.4) suffices to specify the solution across the jump, or if there are further "entropy conditions" that must be satisfied. To answer this, we shall consider (2.2) in a special case, when the ambient field is constant. We define an appropriate boundary value problem for the system, and we show that there exists a unique generalized solution of the boundary value problem. As a consequence, a solution that is smooth in the subregions Ω_1 and Ω_2, and satisfies (2.4), is uniquely determined by its boundary conditions. So, there is no further "entropy condition" that must be satisfied. Finally, we establish the existence of solutions with a jump discontinuity on a given streamline Γ of the ambient field. Thus, discontinuous solutions really do occur.

We consider the particular case of (2.1) for which the ambient flow field satisfies $U(x, y) \equiv U_0 \neq 0$, $V(x, y) \equiv 0$, $P(x, y) \equiv P_0$, $\Theta(x, y) \equiv \Theta_0$. We let $\rho = \rho_0$, $\rho_1 = \rho_{10}$, $\rho_2 = \rho_{20}$, $\varepsilon_1 = \varepsilon_{10}$, $\varepsilon_2 = \varepsilon_{20}$, when evaluated at $P = P_0$, $\Theta = \Theta_0$. In this case, the lower order terms in (2.2) vanish, and (2.2) becomes

$$L_1 q \equiv -(2\mu + \lambda)u_{xx} - \mu u_{yy} - (\mu + \lambda)v_{xy} + \rho_0 U_0 u_x + p_x = 0, \tag{3.1a}$$

$$L_2 q \equiv -(\mu + \lambda)u_{xy} - \mu v_{xx} - (2\mu + \lambda)v_{yy} + \rho_0 U_0 v_x + p_y = 0, \tag{3.1b}$$

$$L_3 q \equiv \rho_0 u_x + \rho_0 v_y + \rho_{10} U_0 p_x + \rho_{20} U_{20}\theta_x = 0, \tag{3.1c}$$

$$L_4 q \equiv P_0 u_x + P_0 v_y + \rho_0 \varepsilon_{10} U_0 p_x - \kappa\Delta\theta + \rho_0 \varepsilon_{20} U_0 \theta_x = 0. \tag{3.1d}$$

For this problem, the possible curves of discontinuity are the lines $y = \text{constant}$. If $y = y^*$ is a line of discontinuity of the solution, the jump conditions (2.4) give

$$\delta u_y(x, y^*) = 0, \qquad (2\mu + \lambda)\delta v_y(x, y^*) = -\delta p(x, y^*).$$

Equation (2.4a) gives $\delta u_x(x, y^*) = 0$. Hence the second jump condition gives $\delta\sigma_{22} = \delta[\lambda u_x + (2\mu + \lambda)v_y] = -\delta p$, so the total normal stress is continuous across Γ.

It is important to clarify a potential misunderstanding concerning the system (3.1). The ambient field is a state of uniform motion, and thus may be obtained from the state of zero flow by a simple change of dependent variable. Nevertheless, the linearized system (3.1) is not related in a simple way to the corresponding system that is obtained by linearizing about a state of zero flow. An intuitive explanation of this arises from the fact that the linearized equations are satisfied by a small perturbation in the flow field. The effect of a nonconstant perturbation on a state of uniform motion is not related in a simple way to the effect of a nonconstant perturbation on the state of zero

motion. Some differences in the two linearized systems may also be seen from the results of the preceding section. As is suggested by the proof of Lemma 1, the system (2.2) is, in fact, elliptic at stagnant points of the ambient flow. Also, the curve Γ of possible discontinuities provided by Theorem 1 is meaningful only at nonstagnant points of the ambient field.

We shall consider the system (3.1) in the rectangle Ω defined by $0<x<1$, $-1<y<1$, with boundary $\partial\Omega$. In [3] it is shown that, for a time-dependent problem, it is appropriate to specify u, v and θ on $\partial\Omega$, and to specify p on the "incoming" portion of the boundary, that is, on the line $x=0$, $-1<y<1$. Thus, our boundary value problem consists of (3.1) with the boundary conditions

$$\begin{aligned} &u,\ v,\ \theta \quad \text{specified on } \partial\Omega, \\ &p(0,y) \quad \text{specified}, \qquad -1<y<1. \end{aligned} \tag{3.2}$$

We shall show that the problem (3.1), (3.2) has a unique solution. Since the numerical values of ρ_0, U_0, P_0, ρ_{10}, ρ_{20}, ε_{10}, and ε_{20} do not affect this analysis, we shall, for ease of notation, take all these values to be 1. It is convenient, for the analysis, to cast the problem into a different form. Subtracting (3.1d) from (3.1c), we obtain

$$-\kappa\Delta\theta=0. \tag{3.1e}$$

We solve this equation with the boundary conditions given by (3.2) to obtain $\theta(x,y)$. We next select functions $\bar{u}(x,y)$, $\bar{v}(x,y)$ which take on the same boundary values as u, v, and we define $\bar{p}(x,y)$ by $\bar{p}_x=-\bar{u}_x-\bar{v}_y-\theta_x$, $\bar{p}(0,y)=p(0,y)$. Considering as unknowns the functions $u-\bar{u}$, $v-\bar{v}$ and $p-\bar{p}$, and renaming these unknowns u, v, p, we are led to the problem

$$-(2\mu+\lambda)u_{xx}-\mu u_{yy}-(\mu+\lambda)v_{xy}+u_x+p_x=f, \tag{3.4a}$$

$$-(\mu+\lambda)u_{xy}-\mu v_{xx}-(2\mu+\lambda)v_{yy}+v_x+p_y=g, \tag{3.4b}$$

$$u_x+v_y+p_x=0 \tag{3.4c}$$

with boundary conditions

$$\begin{aligned} &u,\ v=0 \qquad \text{on } \partial\Omega, \\ &p(0,y)=0, \qquad -1<y<1. \end{aligned} \tag{3.5}$$

To analyze the problem (3.4), (3.5), we shall introduce some Hilbert spaces, define an appropriate bilinear form, and use the Lax-Milgram lemma—techniques that are familiar in the analysis of linear problems. We shall use the Hilbert space $L_2=L_2(\Omega)$, of square integrable functions, the space $H_0^1=H_0^1(\Omega)$ of functions $z(x,y)$ which vanish on $\partial\Omega$ and for which $\|z\|_1^2=\iint\{|\nabla z|^2+z^2\}\,dx\,dy<\infty$, and the dual space $H^{-1}=H^{-1}(\Omega)$. H^{-1} is defined as the set of distributions ϕ such that for some constant c and for $z\in C_0^\infty(\Omega)$, $|\phi(x)|\leqq c\|z\|_1$. The smallest such constant c is the norm $\|\phi\|_{-1}$. Equipped with this norm, and with a corresponding inner product, H^{-1} is a Hilbert space and may be regarded as the space of linear functionals on H_0^1. These spaces satisfy the inclusions $C_0^\infty(\Omega)\subset H_0^1\subset L_2\subset H^{-1}$, and $C_0^\infty(\Omega)$ is dense in each of these spaces in the appropriate norm. We shall also need the spaces $\underline{H}=L_2\times L_2$ and $\underline{V}=H_0^1\times H_0^1$ of pairs of functions on Ω.

Let $\xi=(u,v)\in\underline{V}$ be a pair of functions. Then u_x, $v_y\in L_2$. We define $S\xi=-u_x-v_y$, so $S:\underline{V}\to L_2$ is a bounded map. With ξ given, define $p(x,y)$ by

$$p(x,y)=\int_0^x (S\xi)(s,y)\,ds.$$

Then $p \in L_2$, so $p_y \in H^{-1}$. We define $T\xi = p_y$, so $T: \underline{V} \to H^{-1}$ is a bounded map. We then replace (3.4), (3.5) with the problem: given $(f, g) \in \underline{H}$, find $\xi = (u, v) \in \underline{V}$ so that

$$-(2\mu+\lambda)u_{xx} - \mu u_{yy} - (\mu+\lambda)v_{xy} + u_x + S\xi = f, \tag{3.6a}$$

$$-(\mu+\lambda)u_{xy} - \mu v_{xx} - (2\mu+\lambda)v_{yy} + v_x + T\xi = g. \tag{3.6b}$$

The following theorem establishes the unique solvability of (3.6).

THEOREM 2. *Suppose $\mu > 0$, $3\mu + 2\lambda > 0$. Then there is a constant $c > 0$ such that, given $(f, g) \in \underline{H}$, there exists a unique $\xi = (u, v) \in \underline{V}$ such that ξ solves* (3.6). *Also,*

$$\|\xi\|_{\underline{V}} \leqq c\|(f, g)\|_{\underline{H}}.$$

Proof. Let $\xi = (u, v) \in \underline{V}$, $\eta = (w, z) \in \underline{V}$, and define the bilinear form

$$a(\xi, \eta) = \iint \{(2\mu+\lambda)(u_x w_x + v_y z_y) + \mu(u_y w_y + v_x z_x) + (\mu+\lambda)v_y w_x + (\mu+\lambda)u_x z_y + w u_x + z v_x + w S\xi + z T\xi\}\, dx\, dy.$$

It is easily seen that $a(\xi, \eta)$ is well defined, and that a is a bounded bilinear form on $\underline{V} \times \underline{V}$. Furthermore, $\xi \in \underline{V}$ solves (3.6) if and only if

$$a(\xi, \eta) = \iint [fw + gz]\, dx\, dy, \qquad \eta \in \underline{V}.$$

Next we set $\eta = \xi$. The leading set of terms in $a(\xi, \xi)$ comprise a quadratic form in u_x, u_y, v_x, v_y. A computation shows that this quadratic form is positive definite provided that $\mu(3\mu + 2\lambda) > 0$. Since u and v vanish on $\partial\Omega$, $\iint u u_x\, dx\, dy = \iint v v_y\, dx\, dy = 0$. Set $S\xi = p_x$, $T\xi = p_y$. If u and v lie in $C_0^\infty(\Omega)$, then p is a smooth function, so

$$\iint u S\xi\, dx\, dy = \iint u p_x\, dx\, dy = -\iint u_x p\, dx\, dy.$$

Similarly,

$$\iint v T\xi\, dx\, dy = -\iint v_y p\, dx\, dy.$$

Hence

$$\begin{aligned}\iint [uS\xi + vT\xi]\, dx\, dy &= -\iint (u_x + v_y)p\, dx\, dy \\ &= \iint p p_x\, dx\, dy \\ &= \frac{1}{2}\int_{-1}^{1} p(1, y)^2\, dy \geqq 0,\end{aligned}$$

where the last equality holds because p vanishes on $x = 0$. Hence, using the density of C_0^∞ in $\underline{V}$, for any $\xi \in V$,

$$\iint [uS\xi + vT\xi]\, dx\, dy \geqq 0.$$

Using these facts, we find that

$$a(\xi, \xi) \geqq c \iint [|\nabla u|^2 + |\nabla v|^2]\, dx\, dy,$$

so there is a constant $\alpha > 0$ such that

$$\alpha\|\xi\|_V^2 \leqq a(\xi, \xi), \qquad \xi \in \underline{V}. \tag{3.7}$$

From the Lax-Milgram lemma (see, e.g., [6, III.7]) given $(f, g) \in \underline{H}$, there is a unique $\xi \in \underline{V}$ such that, for any $\eta \in \underline{V}$,

$$a(\xi, \eta) = \int\int [fw + gz]\, dx\, dy.$$

The asserted inequality is a direct consequence of (3.7), and the proof is complete.

We now construct a generalized solution to the equations (3.1) which is not smooth on the line $y = 0$. For simplicity, we shall carry out the construction in the case $\lambda = -\mu$. To carry out the construction, we shall consider the equations in the half plane $x > 0$ instead of in Ω, and we shall seek a solution with $\theta \equiv 0$. We formally take the Fourier transform of the system (3.1a), (3.1b), (3.1c) with respect to y. Setting

$$\tilde{u}(x, t) = \int u(x, y) \exp(-iyt)\, dy,$$

and similarly for $\tilde{v}, \tilde{p}$, we are led to the system

$$-\mu\tilde{u}_{xx} + \mu t^2\tilde{u} + \rho_0 U_0 \tilde{u}_x + \tilde{p}_x = 0, \tag{3.8a}$$

$$-\mu\tilde{v}_{xx} + \mu t^2\tilde{v} + \rho_0 U_0 \tilde{v}_x + it\tilde{p} = 0, \tag{3.8b}$$

$$a\tilde{u}_x + ita\tilde{v} + \tilde{p}_x = 0, \tag{3.8c}$$

where we have set $a = \rho_0/\rho_{10} U_0$.

To solve this system, we define

$$\tilde{V}(x, t) = \int_0^x \tilde{v}(s, t)\, ds.$$

Upon integrating (3.8c), we obtain

$$\tilde{p}(x, t) = -a\tilde{u}(x, t) - ita\tilde{V}(x, t) + \tilde{g}(t), \tag{3.9}$$

where we have set $\tilde{g}(t) = \tilde{p}(0, t) + a\tilde{u}(0, t)$. Using (3.9) to eliminate $\tilde{p}$ from (3.8a), (3.8b), we obtain

$$-\mu\tilde{u}_{xx} + b\tilde{u}_x + \mu t^2\tilde{u} - ita\tilde{V}_x = 0, \tag{3.10a}$$

$$-\mu\tilde{V}_{xxx} + (a + b)\tilde{V}_{xx} + \mu t^2\tilde{V}_x + at^2\tilde{V} - iat\tilde{u} = -it\tilde{g}(t), \tag{3.10b}$$

where we have set $b = \rho_0 U_0 - a$.

To solve (3.10), we consider the homogeneous linear system

$$[-\mu r^2 + br + \mu t^2]\alpha - iatr\beta = 0, \tag{3.11a}$$

$$-iat\alpha + [-\mu r^3 + (a + b)r^2 + \mu t^2 r + at^2]\beta = 0. \tag{3.11b}$$

The determinant of the system (3.11a), (3.11b) is a polynomial $q(r, t)$ which is of degree 5 in r. The roots $r(t)$ of this polynomial are important in our analysis. As $|t| \to \infty$, four of the roots tend to $\pm\infty$, whereas the fifth root tends to $-1/k$, where

$$k = \mu U_0 \rho_{10}/\rho_0.$$

The quantity k, which has the dimensions of length, plays an important role in the propagation of the discontinuities. As we will see, the dimensionless quantity x/k gives the rate of decay of the strength of the discontinuity along the streamline of the ambient flow. The following lemma gives the basic facts concerning the system (3.11) and the quantity k.

LEMMA 2. *There is a constant $d>0$, and piecewise continuous, complex-valued functions $\alpha(t)$, $\beta(t)$, $r(t)$, defined for all real t, such that for each real t, $\beta(t)\neq 0$ and* (3.11a), (3.11b) *is satisfied. As $|t|\to+\infty$, $\lim r(t)=-1/k$ and for $|t|\geqq 1$,*

$$|\alpha(t)|\leqq d/t, \quad 1\leqq d|\beta(t)|, \quad |r(t)+k^{-1}|\leqq d/t. \tag{3.12a}$$

For $|t|\leqq 1$,

$$|\alpha(t)|\leqq d, \quad 1\leqq d|\beta(t)|, \quad |r(t)|\leqq d. \tag{3.12b}$$

Proof. The determinant of the system (3.11a), (3.11b) is the quintic polynomial

$$\begin{aligned} q(r,t) &= \mu^2 r^5-\mu(a+2b)r^4-(2\mu^2t^2-ab-b^2)r^3+2b\mu t^2r^2 \\ &\quad + t^2(a^2+ab+\mu^2t^2)r+a\mu t^4 \\ &= (\mu r^2-(a+b)r-\mu t^2)(\mu r^3-br^2-\mu t^2r-at^2). \end{aligned}$$

We have $q(r,0)=\mu^2r^5-\mu(a+2b)r^4+(ab+b^2)r^3$. Hence $q(r,0)=0$ has roots $r=0$, with multiplicity at least three, and two other roots, $r=(a+b)/\mu$ and $r=b/\mu$. Since $a+b=\rho_0U_0\neq 0$, the root $r=\rho_0U_0/\mu$ of $q(r,0)$ is nonzero. Set $\varepsilon=1/t$. Then for ε near 0,

$$\varepsilon^4 q(r,t)=\mu^2 r+a\mu+O(\varepsilon),$$

so for $|t|$ large, $q(r,t)=0$ has a real root r that is close to $-a/\mu=-1/k$, and satisfies, for some $a>0$,

$$|r+1/k|\leqq d/|t|, \qquad |t|\geqq 1.$$

Also, setting $\chi=\varepsilon r$,

$$\varepsilon^5 q(r,t)=\mu^2\chi(\chi-1)^2(\chi+1)^2-\varepsilon\mu(\chi^2-1)((a+2b)\chi^2+a)+\varepsilon^2(a+b)\chi(b\chi^2+a),$$

so for $|t|$ large, $q(r,t)$ has four roots, λ, such that $\lambda\approx t,\ t,\ -t,\ -t$.

We now pick a piecewise continuous family of roots $r(t)$ of $q(r,t)$ so that $r(t)\to -1/k$ as $|t|\to+\infty$ and $r(t)\to\rho_0U_0/\mu$ as $t\to 0$. Having chosen $r(t)$, we define $[\alpha(t),\ \beta(t)]$ to be a piecewise continuous solution of (3.1) with $|\alpha(t)|^2+|\beta(t)|^2=1$. Then $|\alpha(t)|\leqq 1$, $|\beta(t)|\leqq 1$ for all t. Multiplying the first equation of (3.11) by ε^2 and taking the limit as $|t|\to+\infty$, we obtain $\lim\alpha(t)=0$. A perturbation argument shows that $|\alpha(t)|\leqq c/|t|$ for $|r|>1$. Since $\lim|\beta(t)|=1$, this proves (3.12a). Finally, we show that $\beta(t)\neq 0$ for all t. Suppose, to the contrary, that $\beta(t^*)=0$ for some real t^*. Then $|\alpha(t^*)|=1$, so from (3.11b), $t^*=0$. From (3.11a), either $r(0)=0$ or $r(0)=b/\mu$. This contradicts our choice of $r(t)$ for t near 0, and the proof is complete.

Using the functions $\alpha(t)$, $\beta(t)$, $r(t)$, we define functions $\tilde u(x,t)$, $\tilde v(x,t)$, $\tilde p(x,t)$ by

$$\tilde u(x,t)=\frac{i\tilde g(t)\alpha(t)}{at\beta(t)}\,e^{r(t)x}, \tag{3.13a}$$

$$\tilde v(x,t)=\frac{i\tilde g(t)r(t)}{at}\,e^{r(t)x}, \tag{3.13b}$$

$$\tilde p(x,t)=\tilde g(t)e^{r(t)x}-\frac{i\alpha(t)\tilde g(t)}{t\beta(t)}\,e^{r(t)x}. \tag{3.13c}$$

It may be verified by inspection that $\tilde u$, $\tilde v$, $\tilde p$ satisfy (3.8). Define the function $\tilde g(t)$ by $\tilde g=\tilde g_1-\tilde g_2$, where

$$\tilde g_1(t)=\frac{-1}{t+i},$$

and where $\tilde{g}_2(t)$ is a smooth function of compact support, chosen so that $\tilde{g}_1(0) = \tilde{g}_2(0)$. Then $\tilde{g}(t) = O(t)$ for t near 0, so since $\beta(t)$ is bounded away from 0, the formulas (3.13) have meaning for all real t. We note some properties of $\tilde{g}(t)$.

LEMMA 3. *The function $\tilde{g} \in L_2(R^1)$ and satisfies $|\tilde{g}(t)| \leqq c|t|^{-1}$ for $|t| \geqq 1$. The inverse Fourier transform $g(y) \in L_2(R^1)$ and g is continuous in $(-\infty, 0)$ and $(0, \infty)$ and satisfies $g(+0) - g(-0) = 1$.*

Proof. The inverse Fourier transform of $\tilde{g}_1$ is

$$g_1(y) = \begin{cases} e^{-y}, & y > 0, \\ 0, & y < 0, \end{cases}$$

and the inverse Fourier transform of $\tilde{g}_2$ is a C^∞ function that is rapidly decreasing at ∞. The lemma follows from these facts.

THEOREM 3. *The functions $\tilde{u}$, $\tilde{v}$ and $\tilde{p}$ are, for fixed x, in $H^1(R^1)$, $H^1(R^1)$, $L_2(R^1)$, respectively. The inverse Fourier transforms, u, v, p, are well defined, and $u, v \in H^1(S)$, $p \in L_2(S)$, on each strip $S = \{(x, y): 0 < x < x^*\}$. The functions u, v, p are a weak solution of (3.1). The function $p(x, y)$ is, for fixed x, continuous in y for $y \neq 0$, and has a jump discontinuity at $y = 0$.*

Proof. Using (3.12b) and the construction of $\tilde{g}$, we see that $|\tilde{u}|, |\tilde{v}|$ and $|\tilde{p}|$ are bounded for $|t| \leqq 1$. For $t \geqq 1$, $|\tilde{g}(t)| \leqq c/|t|$ and $|r(t)| \leqq c$, so from (3.12a) and Lemma 2, we have for all t,

$$|\tilde{u}(x, t)| \leqq c/(1+|t|)^3, \qquad |\tilde{u}_x(x, t)| \leqq c/(1+|t|)^3,$$
$$|\tilde{v}(x, t)| \leqq c/(1+|t|)^2, \qquad |\tilde{v}_x(x, t)| \leqq c/(1+|t|)^2,$$
$$|\tilde{p}(x, t)| \leqq c/(1+|t|).$$

Hence for each x,

$$(3.14) \quad \int_{-\infty}^{\infty} [(1+|t|^2)(|\tilde{u}(x, t)|^2 + |\tilde{v}(x, t)|^2) + |\tilde{u}_x(x, t)|^2 + |\tilde{v}_x(x, t)|^2 + |\tilde{p}(x, t)|^2]\, dt \leqq c$$

where the constant c is independent of x if x is bounded. It follows that the inverse transforms are well defined and $u, v \in H^1(S)$, $p \in L_2(S)$, as asserted. To show that u, v, p are weak solutions of (3.1), let $\phi \in C_0^\infty(S)$, and let $\tilde{\phi}(x, t)$ denote the transform of ϕ. Then

$$\iint_S [\mu u_x \phi_x + \mu u_y \phi_y + u_x \phi + p_x \phi)\, dx\, dy = 2\pi \iint_S [\mu \tilde{u} \tilde{\phi}_x + \mu t^2 \tilde{u} \tilde{\phi} + \tilde{u}_x \tilde{\phi} + \tilde{p}_x \tilde{\phi}]\, dx\, dt$$
$$= 0,$$

where the first equality follows from Parseval's formula and the second equality follows from (3.8a). Hence (3.1a) is satisfied in a generalized sense. The other two equations are verified in the same way. It remains to establish the continuity properties of p. From (3.14), for fixed x, $u(x, \cdot) \in H^1(R^1)$. Hence $u(x, y)$ is continuous in y for fixed x. Setting $\tilde{f}(x, t) = \tilde{g}(t) \exp(r(t)x)$, we find from (3.13) that $\tilde{p} = \tilde{f} - a\tilde{u}$, so it suffices to study the continuity properties of the inverse Fourier transform $f(x, y)$ of $\tilde{f}(x, t)$. Define

$$\tilde{f}_1(x, t) = \tilde{g}_1(t) \exp(-x/k),$$
$$\tilde{f}_2(x, t) = \tilde{g}_1(t)[\exp(r(t)x) - \exp(-x/k)],$$
$$\tilde{f}_3(x, t) = \tilde{g}_2(t) \exp(r(t)x).$$

For $|t| \geqq 1$ we use (3.12) to obtain

$$\begin{aligned} |\tilde{f}_2(x, t)| &\leqq |\tilde{g}_1(t)| \cdot |e^{r(t)x} - e^{-x/k}| \\ &\leqq c|t|^{-1}|r(t)x + x/k| \\ &\leqq cxt^{-2}. \end{aligned}$$

Hence, for x in a bounded interval and $-\infty < t < \infty$,

$$|\tilde{f}_2(x, t)| \leqq c(1+|t|)^{-2}.$$

Since $\tilde{g}_2(t)$ has compact support, $\tilde{f}_3(x, t)$ satisfies an even better inequality. Hence, for fixed x the inverse transforms, $f_2(x, y)$ and $f_3(x, y)$ belong to $H^1(R^1)$. Hence, for fixed x, f_2 and f_3 are continuous in y. Since the inverse transform $f_1(x, y) = g_1(y) \exp(-x/k)$, we find that $f_1(x, y)$ is continuous for $y \neq 0$, and $f_1(x, y)$ has a jump discontinuity as $y \to \pm 0$. Hence p has the same properties and the proof is complete.

From the construction of the solution, we see that

$$p(x, y) = g_1(y)\, e^{-x/k} + p_1(x, k), \tag{3.15}$$

where $p_1(x, y)$ is a continuous function. The function $g_1(y)$ has a jump discontinuity at $y = 0$, with a jump of magnitude one. The formula (3.15) shows that the jump is propagated into the region of the flow along the streamline of the ambient flow field that emanates from the point of the discontinuity on the boundary. The factor $\exp(-x/k)$ in (3.15) shows that the magnitude of the jump decays as we move to the interior, and the rate of the decay is measured by the dimensionless quantity x/k. As we move a distance k along the streamline, the magnitude of the jump decreases by a factor of e.

It would be of interest to know if the *nonlinear* system of equations representing steady state viscous barotropic flow has weak solutions with discontinuous pressures. It would also be of interest to know if these discontinuities have been observed experimentally.

4. Conclusions. We have found that the linearized steady state compressible Navier-Stokes equations in two dimensions admit discontinuous solutions. A curve Γ of discontinuities must lie on a streamline of the ambient flow field. The jump conditions for the discontinuity require that the total normal stress to the curve Γ be continuous across Γ, but permit the individual terms that make up this normal stress to be discontinuous. In particular, the pressure may be discontinuous across Γ, but this discontinuity in pressure is balanced by a discontinuity in the normal viscous stress σ_{22} across Γ. The temperature and its first derivatives are continuous across Γ.

In the case when the ambient flow is uniform flow with constant pressure and temperature, a specific discontinuous solution to the linearized system is constructed. The curve of discontinuities in this example emanates from a jump in the pressure on the boundary. The strength of the discontinuity decays as we move into the interior; the jump in pressure is given by the formula

$$\delta p = (\delta\rho)_0 \exp(-x\rho_0/\mu U_0(\partial\rho/\partial P)).$$

Here $(\delta p)_0$ is the jump in pressure at the boundary $x = 0$, and the derivative $\partial\rho/\partial P$ is taken at constant temperature. If the fluid is incompressible, $\partial\rho/\partial P = 0$, and there is no pressure discontinuity in the interior. In the case of an ideal polytropic gas, the dimensionless quantity $x/k = x\rho/\mu U_0(\partial\rho_0/\partial P)$ may be interpreted in terms of the Reynolds number and the Mach number. Setting $\rho = P/R\Theta$, we have $\partial\rho/\partial P = 1/R\Theta = \gamma c^{-2}$, where c is the sound speed and γ is the adiabatic exponent.

Hence in this case,

$$x/k = \frac{x\rho_0 c^2}{\gamma\mu U_0} = \gamma^{-1}M^{-2}\frac{xU_0\rho_0}{\mu} = \gamma^{-1}M^{-2}\,\mathrm{Re},$$

where we have set $M = U_0/c$ = the Mach number of the ambient flow and $\mathrm{Re} = xU_0\rho_0/\mu$ = the Reynolds number. If we consider air at room temperature and atmospheric pressure, the values $c \approx 1100$ ft/sec, $\rho_0 \approx 0.075$ $\mathrm{lb}_m/\mathrm{ft}^3$, $\mu = 1.22\times 10^{-5}$ $\mathrm{lb}_m/\mathrm{ft}$ sec may be found in Whitaker [5]. Hence, retaining $U_0/c = M$, the Mach number, we find that

$$k = \frac{x}{M}\cdot\frac{\rho_0 c}{\gamma\mu} \approx \frac{x}{M}\cdot 4.8\times 10^6,$$

where the distance x is measured in feet. For example, if the ambient flow is at the speed of sound, $M = 1$, the jump in pressure is reduced by a factor of $\exp(-4.8\times 10^6)$ at a distance of one foot from the boundary of the region. This may explain why these discontinuities do not seem to have been discussed in the literature. The pressure jump would decay more slowly with large viscosity.

Acknowledgments. We thank Professors A. Faller, H. Glaz, and T.-P. Liu and Dr. T.-F. Zien for some interesting conversations on this paper.

REFERENCES

[1] H. Beirao de Veiga, *Stationary motions and incompressible limit for compressible viscous fluids*, MRC Report 2883, November 1985; Houston J. Math., to appear.

[2] A. Valli, *On the existence of stationary solutions to compressible Navier-Stokes equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 99-113.

[3] G. Geymonat and P. Leyland, *Transport and propagation of a compressible perturbation through a flow in a bounded region*, to appear.

[4] D. Hoff and T.-P. Liu, to appear.

[5] S. Whitaker, *Introduction to Fluid Mechanics.* Prentice-Hall, Englewood Cliffs, NJ, 1968.

[6] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, 1966.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

006

ON UNIQUENESS OF AXISYMMETRIC DEFORMATIONS OF ELASTIC PLATES AND SHELLS*

HUBERTUS J. WEINITSCHKE†

Abstract. Finite axisymmetric deformations of thin shells of revolution are considered for problems where the radial membrane stress is nonnegative. It is rigorously proved that the solutions of the relevant boundary value problems for both closed and open (doubly connected) shallow shells subjected to arbitrary normal surface load and various edge conditions are unique. This result is shown to hold also for a restricted class of boundary value problems for nonshallow shells.

Key words. geometrically nonlinear deformation of elastic plates and shells, uniqueness of tensile solution in shells, axisymmetric deformation in shells of revolution

AMS(MOS) subject classifications. 73C15, 73C50, 73K10, 73K15, 73L99

1. Introduction. We consider axisymmetric finite deformations of thin elastic shells of revolution. The most frequently encountered geometries in the applications are: a closed shell with no other boundary than the two outer surfaces, a shell closed at the apex, also called a regular dome having a circular boundary with some kind of edge support, or a (doubly connected) ring shell with an additional boundary which may be described as a central opening around the apex. When loads are applied to the surface and to the edges, we generally expect, on intuitive grounds and experience, *uniqueness* of solutions when the stresses throughout the shell are predominantly tensile, for instance if a closed shell is subjected to internal pressure. But shells may of course buckle under suitable loads, usually caused by compressive stresses, and in this case uniqueness can be expected only for sufficiently small loads.

The purpose of this paper is to prove some general uniqueness theorems by elementary methods of classical analysis. These theorems will confirm the intuitive conclusions indicated above. Basically, we assume that the meridional stress resultant N_s is positive throughout the shell, except at the edges (where $N_s = 0$ is permitted), but no assumptions are made on the sign of the circumferential stress resultant N_θ and on the stresses due to the bending moments M_s and M_θ. It turns out that $N_s \geqq 0$ alone implies uniqueness of axisymmetric solutions for a large class of shell problems. More precisely, this type of uniqueness holds for arbitrary shallow shells of revolution, open or closed at the apex, within the framework of small finite deflection theory. Here, N_s is essentially given by the radial stress resultant N_r. In the case of nonshallow shells, we prove uniqueness for a restricted class of boundary value problems in the sense that the boundary conditions for the relevant dependent variables are assumed to be linear. We suspect that uniqueness of axisymmetric solutions with $N_s \geqq 0$ holds true also in the general finite deflection theory for shells of revolution [11], that is, under any given surface load and a general class of physically meaningful edge conditions (linear or nonlinear) there cannot be more than one solution with N_s nonnegative. On the other hand, it is well known that $N_s \geqq 0$ does not imply global uniqueness, as shells of revolution may buckle asymmetrically when N_θ is compressive.

* Received by the editors April 23, 1986; accepted for publication (in revised form) May 20, 1987.

† Institute of Applied Mathematics, University of Erlangen-Nürnberg, 8520 Erlangen, West Germany. This work was done while the author was visiting at the Institute of Applied Mathematics, University of British Columbia, Vancouver, Canada. This work was supported in part by Natural Sciences and Engineering Research Council of Canada grant A-5201.

The problem of the *existence* of solutions for shallow spherical shells, closed at the apex, subjected to normal pressure and various edge conditions, has been resolved by Wagner [15], using methods of the calculus of variations. A different proof for the special case of a circular plate was later given by Dickey [2]. The results obtained in [15] imply a uniqueness proof for circular plate boundary value problems as shown by Reiss [10]. Here we give an elementary uniqueness proof for arbitrary shallow shells without using Wagner's result. In the special case of circular plates, our results specialize to those in [10]. Furthermore, we obtain new uniqueness results for ring shells, including annular plates as a special case. Apparently, the von Kármán equations for annular plates have not been analyzed in previous work with regard to the mathematical questions of existence and uniqueness of solutions. For annular flat membrane problems, these questions have recently been treated by the author and Grabmüller [19], [5] (see also recent work by Grabmüller and Novak [3], [4]).

2. Basic equations, shallow shells. Our starting point is the basic equations of shallow shell theory in the form given by E. Reissner [11], [12]. Restricting ourselves to axisymmetric stresses and displacements in shells of revolution, the equations can be written in terms of dimensionless variables as follows [17]:

$$Lf = -Z(x)g + fg + 2\gamma R(x, \varepsilon), \quad Lg = Z(x)f - \tfrac{1}{2}f^2, \quad \varepsilon < x < 1 \tag{2.1}$$

where

$$L = \frac{d^2}{dx^2} + \frac{3}{x}\frac{d}{dx}, \qquad R(x, \varepsilon) = \frac{2}{x^2}\int_{\varepsilon}^{x} \xi \bar{p}(\xi)\, d\xi.$$

f relates to an angular deflection, g to the radial stress σ_r, $Z(x)$ to the geometry of the undeformed middle surface, and $x = r/a$, $\varepsilon = r_i/a$, where a is the base radius of the outer edge and r_i is the radius of the central opening, called the inner edge. The polarly symmetric pressure $p(r)$ is scaled to $\bar{p} = p/p_0$, where $p_0 = \max |p(r)|$. Hence, γ is a measure of the intensity of the applied load, $\gamma \sim p_0$, and γ is positive for external pressure. In the special case of a spherical shell we have $Z(x) = \mu^2$, where $\mu^2 = 2m_0 H_0/t$, H_0 being the height of the apex above the base plane through $r = a$, t the shell thickness and $m_0^2 = 12(1-\nu^2)$, ν = Poisson's ratio.

At the outer edge $r = a$ we prescribe

$$g(1) = S \quad \text{or} \quad g'(1) + (1-\nu)g(1) = H \tag{2.2}$$

where S is a given dimensionless radial traction and H is a dimensionless radial displacement. For a clamped or a moment-supported outer edge we have, in addition to (2.2),

$$f(1) = 0 \quad \text{or} \quad f'(1) + (1+\nu)f(1) = M, \tag{2.3}$$

respectively, where M is a prescribed dimensionless radial edge moment.

If the shell is closed at the apex $r = 0$, we formally set $\varepsilon = 0$ in (2.1) and assume the usual regularity conditions

$$f'(0) = 0, \qquad g'(0) = 0. \tag{2.4}$$

Equations (2.1)-(2.4) then constitute the boundary value problem (BVP) for a closed (dome) shell, denoted as Problem I. Selecting one boundary condition from (2.2) and (2.3) we obtain the four different BVP's (C, S), (M, S), (C, H), and (M, H), with obvious notation and C referring to the clamped edge condition $f(1) = 0$, although it may be replaced by $f(1) = C$ if the angle of rotation is prescribed.

When the shell has a circular opening at the apex, we replace (2.4) by similar boundary conditions at the inner edge $r = r_i$, expressed in terms of f and g by

$$f(\varepsilon) = c \quad \text{or} \quad \varepsilon f'(\varepsilon) + (1+\nu) f(\varepsilon) = m \tag{2.5}$$

and

$$g(\varepsilon) = s \quad \text{or} \quad \varepsilon g'(\varepsilon) + (1-\nu) g(\varepsilon) = h. \tag{2.6}$$

The physical significance is analogous to that of conditions (2.2) and (2.3). In particular, the case of a "free edge" $g(\varepsilon) = \varepsilon f'(\varepsilon) + (1+\nu) f(\varepsilon) = 0$ is contained in (2.5), (2.6). Equations (2.1)-(2.3), (2.5), and (2.6) for $\varepsilon > 0$ define the boundary value problem for shells with a central hole, which will henceforth be denoted as Problem II. The special case of a uniform pressure is given by $R(x, \varepsilon) = 1 - (\varepsilon/x)^2$. Selecting one boundary condition from each of the sets (2.2), (2.3), (2.5), and (2.6) we obtain a total of 16 BVP's denoted by 4-tuples such as $(c, s; M, H)$, in obvious generalization of the notation for Problem I.

3. Uniqueness of tensile solutions of Problem I. Let (f_1, g_1) and (f_2, g_2) be two solutions of Problem I or II, then we have from (2.1), setting $v = f_1 - f_2$, $w = g_1 - g_2$,

$$\begin{aligned} Lv &= -Zw + f_1 g_1 - f_2 g_2 = -Zw + g_1 v + f_2 w, \\ Lw &= Zv - \tfrac{1}{2}(f_1^2 - f_2^2) = -\tfrac{1}{2}(f_1 + f_2 - 2Z) v, \end{aligned}$$

with homogeneous boundary conditions corresponding to (2.2)-(2.6) for v and w. Now we integrate $vLv + wLw$ after multiplying by x^3

$$\begin{aligned} \int_\varepsilon^1 (vLv + wLw) x^3 \, dx &= \int_\varepsilon^1 [v(x^3 v')' + w(x^3 w')'] \, dx \\ &= \int_\varepsilon^1 \left[-Zvw + g_1 v^2 - f_2 vw - \frac{1}{2} vw(f_1 + f_2 - 2Z) \right] x^3 \, dx. \end{aligned} \tag{3.1}$$

Integrating by parts and simplifying the second line of (3.1), we obtain

$$-\int_\varepsilon^1 x^3 (v'^2 + w'^2) \, dx + B_\varepsilon + B_1 = \frac{1}{2} \int_\varepsilon^1 x^3 v^2 (g_1 + g_2) \, dx. \tag{3.2}$$

In Problem I, $\varepsilon = 0$ and $B_0 = 0$ by (2.4), while $B_1 = v(1)v'(1) + w(1)w'(1)$, which is therefore given by

$$B_1 = -\alpha(1+\nu) v^2(1) - \beta(1-\nu) w^2(1) \leqq 0, \tag{3.3}$$

where α and β take on the values 0 or 1, depending on which combination of boundary conditions from (2.2) and (2.3) is selected. Hence, the left-hand side of (3.2) is always nonpositive, while the right-hand side is nonnegative, provided $g_i(x) \geqq 0$, $i = 1, 2$.

DEFINITION 1. A solution (f, g) of Problem I or II is called *tensile*, if $g(x) \geqq 0$, which is equivalent to $\sigma_r \geqq 0$.

Note that no conditions (other than smoothness) are imposed on the remaining stress components or bending moments for a solution to be tensile in our terminology.

THEOREM 3.1. *Tensile solutions* $(f(x), g(x))$ *of Problem* I *are unique.*

Proof. Since (3.2) implies $v' = w' \equiv 0$, we have $v(x) = v_0 = \text{const.}$ and $w(x) = w_0 = \text{const.}$ For $\alpha = \beta = 1$, that is BVP (M, H), (3.3) shows that $B_1 < 0$ unless $v(1) = v_0 = w(1) = w_0 = 0$; hence $v = w \equiv 0$. In BVP (M, S), $\beta = 0$, $\alpha = 1$, we conclude $v_0 = 0$ from (3.3), implying $v \equiv 0$, and since $g_1(1) = g_2(1) = S$, we also have $w_0 = 0$; hence $w \equiv 0$. A similar argument shows $v = w \equiv 0$ in BVP (C, H), where $\alpha = 0$, $\beta = 1$. Finally, $v(1) = w(1) = 0$ in BVP (C, S), which completes the proof.

In the case of a circular plate $Z=0$; hence the second equation of (2.1) implies

$$x^3 g'(x) = -\frac{1}{2}\int_0^x t^3 f^2(t)\,dt \leqq 0. \tag{3.4}$$

If $g(1)=S\geqq 0$, this yields $g(x)\geqq S$, as g is monotone nonincreasing. If $g'(1)+(1-\nu)g(1)=H$, with $H\geqq 0$ (to exclude buckling), $g'(1)\leqq 0$ from (3.4) implies $g(1)\geqq 0$, whence $g(x)\geqq g(1)$. Thus the solutions (f, g) are tensile and we conclude the following.

THEOREM 3.2. *The solutions of the circular plate BVP's* (C, S), (M, S), (C, H), *and* (M, H) *are unique, provided* $S\geqq 0$ *or* $H\geqq 0$.

The same result can be expected to hold for sufficiently small $Z(x)$. On the other hand, axisymmetric buckling of circular plates is known to occur in the range $S<0$ or $H<0$. A slightly less general version of Theorem 3.2 was proved in a different way by Reiss [10], on the basis of results of Wagner [15].

4. Uniqueness ot tensile solutions of Problem II. It is seen that the method of proving Theorem 3.1 carries over to Problem II only if $B_\varepsilon = -\varepsilon^3(vv'+ww')$ is zero, that is, if the boundary conditions are $f(\varepsilon)=c$ *and* $g(\varepsilon)=s\geqq 0$. If all other cases we have

$$B_\varepsilon = \varepsilon^2[\alpha(1+\nu)v^2(\varepsilon)+\beta(1-\nu)w^2(\varepsilon)]\geqq 0,$$

with $\alpha=1$ or $\beta=1$ or $\alpha=\beta=1$. Hence, we cannot conclude that the left-hand side of (3.2) is nonpositive.

In order to get the proper sign for B_ε, a transformation due to Schwerin [13] was used in earlier work on annular membranes [5]. This transformation can be adapted to our Problem II in several ways. For instance, let

$$z=\frac{x^2-\varepsilon^2}{1-\varepsilon^2},\quad f(x)=\frac{1}{x^2}\bar{f}(z),\quad g(x)=\bar{g}(z). \tag{4.1}$$

The interval $[\varepsilon, 1]$ is mapped onto $[0, 1]$, and any regular tensile solution (f, g) is transformed into such a solution $(\bar{f}, \bar{g})$ and vice versa. The differential equations for Problem II in terms of z, $\bar{f}$ and $\bar{g}$ are

$$\begin{aligned} k_\varepsilon^2\ddot{\bar{f}} &= -Z\bar{g}+2\gamma R+x^{-2}\bar{f}\bar{g}, \\ k_\varepsilon^2 x^2\ddot{\bar{g}}+4k_\varepsilon\dot{\bar{g}} &= Zx^{-2}\bar{f}-\tfrac{1}{2}x^{-4}\bar{f}^2, \end{aligned} \qquad 0<z<1 \tag{4.2}$$

where a dot denotes differentiation with respect to z and $x^2=\varepsilon^2+z(1-\varepsilon^2)$. The boundary conditions at $x=\varepsilon$ become

$$\bar{f}(0)=c\varepsilon^2 \quad \text{or} \quad \dot{\bar{f}}(0)-k_{-\nu}\bar{f}(0)=\bar{m}, \tag{4.3}$$

$$\bar{g}(0)=s \quad \text{or} \quad \dot{\bar{g}}(0)+k_{-\nu}\bar{g}(0)=\bar{h}\varepsilon^{-2} \tag{4.4}$$

and the boundary conditions at $x=1$ are

$$\bar{f}(1)=C \quad \text{or} \quad \dot{\bar{f}}(1)-\varepsilon^2 k_{-\nu}\bar{f}(1)=\bar{M}, \tag{4.5}$$

$$\bar{g}(1)=S \quad \text{or} \quad \dot{\bar{g}}(1)+\varepsilon^2 k_{-\nu}\bar{g}(1)=\bar{H}. \tag{4.6}$$

The constants in (4.2)–(4.6) are defined by

$$k_\varepsilon=\frac{2}{1-\varepsilon^2},\quad k_{\pm\nu}=\frac{1-\varepsilon^2}{2\varepsilon^2}(1\pm\nu),\quad (\bar{m}, \bar{h}, \bar{M}, \bar{H})=k_\varepsilon^{-1}(m, h, M, H). \tag{4.7}$$

Now let $(\bar f_i, \bar g_i)$ $i=1,2$ be two tensile solutions of (4.2)-(4.6), set $v=\bar f_1-\bar f_2$, $w=\bar g_1-\bar g_2$, $Q(z):=[\varepsilon^2+z(1-\varepsilon^2)]^2$ and carry out essentially the same steps which in § 3 led to the identity (3.2). Making use of the relation $x^4\ddot{\bar g}+(4x^2/k_\varepsilon)\dot{\bar g}=(Q\dot{\bar g})^{\cdot}$, the result is

$$-\int_0^1 (Q\dot w^2+\dot v^2)\,dz+B_0+B_1=\frac{1}{8}(1-\varepsilon^2)^2\int_0^1 Q^{-1/2}v^2(\bar g_1+\bar g_2)\,dz \tag{4.8}$$

where

$$B_0:=-v(0)\dot v(0)-\varepsilon^4 w(0)\dot w(0),\qquad B_1:=v(1)\dot v(1)+w(1)\dot w(1). \tag{4.9}$$

It is seen from (4.3) and (4.6) that the signs are such that we can conclude $B_0+B_1\leqq 0$ upon substituting $\dot v(0)$ into (4.9) and taking $w(0)=v(1)=w(1)=0$ or substituting $\dot v(0)$ and $\dot w(1)$ into (4.9) and taking $w(0)=v(1)=0$. Hence, uniqueness for the BVP's $(m, s; C, S)$ and $(m, s; C, H)$ follows as in the proof of Theorem 3.1.

A similar conclusion holds in the dual case where (4.1) is replaced by

$$z=\frac{x^2-\varepsilon^2}{1-\varepsilon^2},\quad f(x)=\bar f(z),\quad g(x)=\frac{1}{x^2}\bar g(z). \tag{4.10}$$

The boundary conditions now transform into

$$\bar f(0)=c \quad\text{or}\quad \dot{\bar f}(0)+k_{+\nu}\bar f(0)=\bar m\varepsilon^{-2}, \tag{4.11}$$

$$\bar g(0)=s\varepsilon^2 \quad\text{or}\quad \dot{\bar g}(0)-k_{+\nu}\bar g(0)=\bar h, \tag{4.12}$$

$$\bar f(1)=C \quad\text{or}\quad \dot{\bar f}(1)+\varepsilon^2 k_{+\nu}\bar f(1)=\bar M, \tag{4.13}$$

$$\bar g(1)=S \quad\text{or}\quad \dot{\bar g}(1)-\varepsilon^2 k_{+\nu}\bar g(1)=\bar H. \tag{4.14}$$

The identity corresponding to (4.8) is easily computed. We obtain

$$-\int_0^1 (Q\dot v^2+\dot w^2)\,dz+B_0+B_1=\frac{1}{8}(1-\varepsilon^2)^2\int_0^1 v^2(\bar g_1+\bar g_2)\,dz, \tag{4.15}$$

where $B_0:=-\varepsilon^2 v(0)\dot v(0)-w(0)\dot w(0)$ and B_1 is defined as in (4.9). Now the signs in (4.12) and (4.13) allow us to conclude $B_0+B_1\leqq 0$ for the BVP's $(c, h; C, S)$ and $(c, h; M, S)$, which implies uniqueness. Since the cases $(c, s; \cdot, \cdot)$ have already been resolved at the beginning of this section, we may summarize the results in the following.

THEOREM 4.1. *Tensile solutions* $(f(x), g(x))$ *of Problem* II *are unique for the following BVP's*:

$$\begin{aligned}&(c, s; C, S),\quad (c, s; M, S),\quad (c, s; C, H),\quad (c, s; M, H),\\ &(m, s; C, S),\quad (m, s; C, H),\quad (c, h; C, S),\quad (c, h; M, S).\end{aligned} \tag{4.16}$$

Obviously, one more case may be settled with the help of the transformation

$$z=\frac{x^2-\varepsilon^2}{1-\varepsilon^2},\quad f(x)=\frac{1}{x^2}\bar f(z),\quad g(x)=\frac{1}{x^2}\bar g(z). \tag{4.17}$$

In that case the boundary conditions are given by (4.3), (4.12), (4.5), and (4.14). Hence, $B_0\leqq 0$ but $B_1\geqq 0$, in the identity corresponding to (4.15), except in the single case $(m, h; C, S)$, where $B_1\leqq 0$. The above method has been discussed here as it may be useful in a variety of other BVP's. However, its scope is limited by the fact that transformations such as (4.1) and (4.10) change the sign in boundary conditions involving first derivatives at both ends of the interval, for example in (4.3) and (4.5), or in (4.12) and (4.14). In Problem II, this means that BVP's of type $(\cdot, h; \cdot, H)$ or $(m, \cdot; M, \cdot)$ cannot be handled by the above technique.

We next present a more flexible approach, in order to resolve the remaining eight cases not covered by (4.16) in Theorem 4.1. Again this method may be useful for a variety of other BVP's. Although its scope is limited in a different way, the method turns out to be general enough to yield uniqueness in all 16 BVP's for Problem II, including the ones settled by Theorem 4.1. However, this is due to the particular structure of both differential equations and boundary conditions of the present problem.

The variables x, f, g are transformed in accordance with (4.17). Let (f_i, g_i), $i=1, 2$ be two tensile solutions of Problem II and consider the weighted differences

$$v(z)=\frac{\bar f_1(z)-\bar f_2(z)}{p(z)}, \quad w(z)=\frac{\bar g_1(z)-\bar g_2(z)}{q(z)}, \quad p, q>0.$$

Subtracting the transformed differential equations we first get

$$k_\varepsilon^2(vp)^{\cdot\cdot}=-x^{-2}Zwq+x^{-4}(\bar f_1\bar g_1-\bar f_2\bar g_2),$$
$$k_\varepsilon^2(wq)^{\cdot\cdot}=x^{-2}Zvp-\tfrac{1}{2}x^{-4}(\bar f_1^2-\bar f_2^2).$$

Multiplying the first equation by pv, the second by qw, adding the resulting equations and integrating, we obtain, after simplifying the right-hand term in the same way as in § 3,

$$k_\varepsilon^2\int_0^1[vp(vp)^{\cdot\cdot}+wq(wq)^{\cdot\cdot}]\,dz=\int_0^1\frac{p^2v^2}{2Q(z)}(\bar g_1+\bar g_2)\,dz$$

where k_ε and $Q(z)$ have been defined before. The choice

$$p=Az+B, \quad q=Dz+E, \quad A, B, D, E=\text{positive constants}$$

implies $p(vp)^{\cdot\cdot}=(p^2\dot v)^{\cdot}$, and integration by parts yields

$$-\int_0^1(p^2\dot v^2+q^2\dot w^2)\,dz+\bar B_0+\bar B_1=\int_0^1\left(\frac{pv}{k_\varepsilon}\right)^2\frac{\bar g_1+\bar g_2}{2Q}\,dz \tag{4.18}$$

where

$$\bar B_0=-p^2v\dot v-q^2w\dot w|_{z=0}, \qquad \bar B_1=p^2v\dot v+q^2w\dot w|_{z=1}.$$

The boundary conditions for $\bar f$ and $\bar g$ are given by (4.3), (4.5), (4.12), and (4.14). Note that $\dot v$, $\dot w$ are now given by

$$\dot v=\frac{1}{p}(\bar f_1-\bar f_2)^{\cdot}-\frac{\dot p}{p^2}(\bar f_1-\bar f_2), \qquad \dot w=\frac{1}{q}(\bar g_1-\bar g_2)^{\cdot}-\frac{\dot q}{q^2}(\bar g_1-\bar g_2).$$

Evaluating the term $p^2v\dot v$ at $z=1$, we find

$$p^2v\dot v=p^2v^2\left[-\frac{\dot p}{p}+\frac{1-\varepsilon^2}{2}(1-\nu)\right]=p^2v^2\left[-\frac{A}{A+B}+\frac{1-\varepsilon^2}{2}(1-\nu)\right].$$

Setting $A=K_1B$, we can choose $K_1>0$ such that

$$\frac{K_1}{1+K_1}>\frac{1-\varepsilon^2}{2}(1-\nu), \tag{4.19a}$$

which makes $p^2(1)v(1)\dot v(1)\leqq 0$. Similarly, with $D=K_2E$ we can choose $K_2>0$ such that

$$\frac{K_2}{1+K_2}>\frac{1-\varepsilon^2}{2}(1+\nu) \tag{4.19b}$$

in order to have $q^2(1)w(1)\dot{w}(1)\leqq 0$. By the same device, we can achieve that $-p^2v\dot{v}$ and/or $-q^2w\dot{w}$ at $z=0$ are negative, that is, $\bar{B}_0\leqq 0$. From (4.3) we find, at $z=0$,

$$p^2v\dot{v}=p^2v^2\left(k_{-\nu}-\frac{\dot{p}}{p}\right)=p^2v^2\left[\frac{1-\varepsilon^2}{2\varepsilon^2}(1-\nu)-\frac{A}{B}\right]$$

and a similar expression for $q^2w\dot{w}$. Clearly the choice

$$\text{(4.20a, b)}\qquad 0<\frac{A}{B}=K_1<\frac{1-\varepsilon^2}{2\varepsilon^2}(1-\nu),\qquad 0<\frac{D}{E}=K_2<\frac{1-\varepsilon^2}{2\varepsilon^2}(1+\nu)$$

will make $\bar{B}_0\leqq 0$. Hence $p(z)>0$ and $q(z)>0$ can be found such that $\bar{B}_0+\bar{B}_1\leqq 0$ in the seven BVP's described by

$$\text{(4.21a, b)}\qquad (c, s; \cdot, \cdot),\qquad (\cdot, \cdot; C, S)$$

and in the two BVP's

$$\text{(4.22a, b)}\qquad (c, h; M, S),\qquad (m, s; C, H).$$

In the cases (4.21a), K_1 or K_2 or both must satisfy (4.19), in the cases (4.21b), they must satisfy (4.20). The BVP's (4.22) require that K_1 and K_2 must satisfy (4.19a) and (4.20b), or (4.19b) and (4.20a), respectively. Incidentally, the BVP's (4.21) and (4.22) are precisely those covered by Theorem 4.1 and the remark following it.

It remains to show that p and q can be chosen such that the cases where both h and H and/or m and M are involved, that is,

$$\text{(4.23)}\qquad (\cdot, h; \cdot, H)\quad \text{and/or}\quad (m, \cdot; M, \cdot),$$

can be resolved. In order to make both $-p^2(0)v(0)\dot{v}(0)$ and $p^2(1)v(1)\dot{v}(1)$ nonpositive, K_1 must satisfy the two inequalities

$$\text{(4.24a, b)}\qquad \frac{K_1}{1+K_1}\geqq\frac{1-\varepsilon^2}{2}(1-\nu)=\varepsilon^2k_{-\nu}\quad\text{and}\quad K_1<\frac{1-\varepsilon^2}{2\varepsilon^2}(1-\nu)=k_{-\nu}.$$

Taking equality in (4.24a), we get $K_1=\varepsilon^2k_{-\nu}/(1-\varepsilon^2k_{-\nu})$, which indeed satisfies (4.24b), because the function

$$\phi_-(\varepsilon)=\frac{\varepsilon^2}{1-\varepsilon^2k_{-\nu}}=\frac{2\varepsilon^2}{1+\nu+(1-\nu)\varepsilon^2}$$

satisfies $0<\phi_-(\varepsilon)<1$ for all ε, ν with $0<\varepsilon<1$, $\nu>-1$. Finally, in order to make both $-q^2(0)w(0)\dot{w}(0)$ and $q^2(1)w(1)\dot{w}(1)$ nonpositive, K_2 is chosen such that it satisfies inequalities obtained from (4.24) by replacing K_1 and $k_{-\nu}$ by K_2 and $k_{+\nu}$, respectively. This can be done since $0<\phi_+(\varepsilon)<1$, with $\phi_+(\varepsilon)=\varepsilon^2/(1-\varepsilon^2k_{+\nu})$. With these values of K_1 and K_2, uniqueness for the seven BVP's (4.23), which include $(m, h; M, H)$, is proved. In summary, we have obtained the analogue of Theorem 3.1, that is, the following.

THEOREM 4.2. *Tensile solutions* $(f(x), g(x))$ *of Problem* II *are unique.*

We remark that it is important that at least one of two inequalities (4.24) is not an equality, because we must show from $v'=w'\equiv 0$ that in all combinations of boundary conditions we actually have $v=w\equiv 0$, as was done in the proof of Theorem 3.1. Obviously, the same type of arguments carry over to Problem II.

5. An application to annular plates. In this section we prove that the solutions of all annular plate BVP's are tensile provided the edge stresses s and S are nonnegative and that any prescribed radial displacements h and H satisfy certain inequalities so

as to exclude buckling. But we shall not insist on the rather strong conditions $h \leqq 0$, $H \geqq 0$. Applying Theorem 4.2, we then get an extension of Theorem 3.2 to annular plate problems.

We transform the second equation of (2.1), with $Z = 0$, according to (4.10), and restate the resulting boundary conditions for $\bar{g}(z)$

$$k_\varepsilon^2 \ddot{\bar{g}}(z) = -\frac{1}{2}\bar{f}(z)^2 \qquad \begin{cases} \bar{g}(0) = s\varepsilon^2 & \text{or} \quad \bar{g}(0) - k_{+\nu}\bar{g}(0) = \bar{h}, \\ \bar{g}(1) = S & \text{or} \quad \dot{\bar{g}}(1) - \varepsilon^2 k_{+\nu}\bar{g}(1) = \bar{H}. \end{cases} \tag{5.1}$$

In the cases $(s, \cdot\,; S, \cdot)$ it follows that $\bar{g}(z)$ satisfies the equation

$$\bar{g}(z) = Az + B + \frac{1}{2k_\varepsilon^2}\int_0^1 G(z, \xi)\bar{f}(\xi)^2\, d\xi, \tag{5.2}$$

where

$$A = S - s\varepsilon^2, \qquad B = s\varepsilon^2, \qquad G(z, \xi) = \begin{cases} (1-\xi)z, & 0 \leqq z \leqq \xi, \\ (1-z)\xi, & \xi \leqq z \leqq 1. \end{cases}$$

G is simply Green's function for $-\ddot{g}$ and the selected boundary conditions. Clearly, for s, S nonnegative we have $\bar{g}(z) \geqq 0$.

The remaining three cases are handled in the same way. Again $\bar{g}(z)$ satisfies (5.2); the quantities A, B and G are summarized in Table 1, where $k = k_{+\nu}$, $k_0 = \varepsilon^2 k/(1-\varepsilon^2 k)$ and G for $\xi \leqq z \leqq 1$ is obtained from $G(z, \xi) = G(\xi, z)$. Obviously, we have $G \geqq 0$. The conditions $A + B \geqq 0$ and $B \geqq 0$ then imply $\bar{g}(z) \leqq 0$. Evaluation of these conditions yields

$$\begin{aligned} &(s, \cdot\,; S, \cdot) && S \geqq 0, \quad s \geqq 0, \\ &(s, \cdot\,; H, \cdot) && H(1-\varepsilon^2) + 2s\varepsilon^2 \geqq 0, \quad s \geqq 0, \\ &(h, \cdot\,; S, \cdot) && S \geqq 0, \quad h(1-\varepsilon^2) \leqq 2S, \\ &(h, \cdot\,; H, \cdot) && H \geqq h\left[1 - \frac{1-\varepsilon^2}{2}(1+\nu)\right], \quad \left[1 + \frac{1-\varepsilon^2}{2\varepsilon^2}(1+\nu)\right]H \geqq h, \end{aligned} \tag{5.3}$$

where (4.7) has been used. We note that the inequalities in (5.3) are identical with those in [5], where they delineate the range of values of s, S, h and H, for which existence and uniqueness of the corresponding annular membrane BVP's was proved. The above results may be summarized in the following.

THEOREM 5.1. *The solutions of all annular plate BVP's are unique, provided that s, S, h and H satisfy the inequalities given in* (5.3), *for the four types of Problem* II, *with no restrictions on the values of c, C, m and M, to be inserted into the slots of the symbols* $(s, \cdot\,; S, \cdot)$, *etc.*

TABLE 1

Case	A	B	$G(z, \xi)$, $\quad 0 \leqq z \leqq \xi$
$(s, \cdot\,; H, \cdot)$	$\dfrac{\bar{H} + \varepsilon^4 sk}{1-\varepsilon^2 k}$	$s\varepsilon^2$	$(1 + k_0\xi)z$
$(h, \cdot\,; S, \cdot)$	$\dfrac{\bar{h} + kS}{1+k}$	$\dfrac{S - \bar{h}}{1+k}$	$(1-\xi)\dfrac{1+kz}{1+k}$
$(h, \cdot\,; H, \cdot)$	$\dfrac{\bar{H} - \varepsilon^2\bar{h}}{1-\varepsilon^2(1+k)}$	$\dfrac{\bar{H} - \bar{h}(1-\varepsilon^2 k)}{k[1-\varepsilon^2(1+k)]}$	$(1 + k_0\xi)\dfrac{1+kz}{k-k_0}$

The conditions contained in (5.3) are satisfied if $H \geqq 0$ and $h \leqq 0$. However, $H < 0$ and $h > 0$ are not excluded. For example, $H < 0$ is admitted in BVP's of the type $(s, \cdot\,; H, \cdot)$ and $h > 0$ is admitted in BVP's of type $(h, \cdot\,; S, \cdot)$. It would be interesting to see how close the bounds in (5.3) are to the critical values where $\bar{g}$ ceases to be nonnegative.

6. Tensile solutions of BVP's for nonshallow shells of revolution. A system of equations governing the axisymmetric deformation of thin elastic shells of revolution undergoing small strains but arbitrarily large rotations was first derived by E. Reissner [11], and later simplified in [12]. This system reduces to a coupled pair of second-order differential equations for the meridional angle of deformation β and stress function ψ (defined below), which generalizes (2.1) to nonshallow shells. Existence of solutions of Reissner's equations for a limited class of shell problems, including ring shells with boundary conditions $\beta = \psi = 0$ at the inner edge, but excluding dome type shells, has been proved via the Leray–Schauder fixed point theorem by Srubshchik [14]. It is well known that these methods cannot be used for answering questions about the uniqueness of solutions.

As Koiter remarked, the nonlinear equations for shells of revolution "have always been (slightly) disfigured by the occurrence of (small) terms with Poisson's ratio ν as a factor" [7]. It has been observed by several authors, including Reissner and Koiter, that these terms never affect the solution significantly within the basic accuracy of first-approximation shell theory. Furthermore, it was shown in [7] and [9] that the terms in question do not appear at all, if the basic equations are derived from the well-founded general intrinsic equations of nonlinear shell theory. In view of these insights and the remarks in [12], we may simplify the basic equations (III) and (IV) in [11] by dropping the small terms multiplied by ν and some terms Reissner himself recognized as negligibly small in [11], in particular the terms containing the vertical stress resultant P_v on the right of equation (IV). The resulting equations (6.1) are given below. In many studies of axisymmetric buckling of hemispherical and complete spherical shells, the simplified equations (6.1) have been used, for example in [8]. Similar equations were used in [1] and [6]. Hence, we may take (6.1) as a realistic model for nonshallow shells of revolution.

Let the shell mid-surface be given in cylindrical coordinates by $r = r(s)$, $z = z(s)$, and let primes denote differentiation with respect to the meridional parameter s, $0 \leqq s_0 \leqq s \leqq s_1$ (for dome shells $s_0 = 0$). We introduce the angle ϕ by $r' = \alpha \cos \phi$ and $z' = \alpha \sin \phi$ where

$$\alpha(s) = [r'(s)^2 + z'(s)^2]^{1/2}.$$

The basic equations in [11], with the simplifications described above, can be written as

$$\beta'' + \frac{(r/\alpha)'}{r/\alpha}\beta' - \left(\frac{r'}{r}\right)^2 \beta = -\frac{\alpha^2}{rD}[\psi \sin \phi - rP_v \cos \phi - \beta(\psi \cos \phi + rP_v \sin \phi)],$$

$$\psi'' + \frac{(r/\alpha)'}{r/\alpha}\psi' - \left(\frac{r'}{r}\right)^2 \psi = \frac{\alpha^2 C}{r}(\beta \sin \phi - \tfrac{1}{2}\beta^2 \cos \phi) + A_1(s) rP_v + A_2(s)(rP_v)' + A_3(s) r^2 p_r + A_4(s)(r^2 p_r)' \tag{6.1}$$

where

$$\psi = rN_r, \qquad rP_v = -\int_{s_0}^{s} rp_v \alpha d\xi,$$

(6.2) N_r, P_v = radial (=horizontal) and vertical stress resultants,
p_r, p_v = radial (=horizontal) and vertical surface load intensities,
$C = Et$ = (constant) stretching stiffness,
$D = Et^3/12(1-\nu^2)$ = (constant) bending stiffness.

The functions $A_i(s)$ are given in terms of r, z, r' and z' [11], but as they disappear when forming $\psi_1 - \psi_2$, their explicit form is irrelevant for deriving uniqueness results. We have retained only quadratic nonlinear terms involving β and ψ, that is, we consider small finite displacements, as in the preceding sections. The fully nonlinear system for arbitrary finite displacements (see equations (I), (II) in [11]) involve terms like $\sin(\beta+\phi)$ and $\cos(\beta+\phi)$ instead of β and β^2 on the right of (6.1). In the shallow shell approximation, $\phi(s)$ is treated as a small quantity, resulting in the approximations $r \sim a_0 s$, where a_0 is a constant, $r'/r \sim 1/s$, $r/\alpha \sim s$, and $\alpha^2 \cos\phi(s) \sim \alpha_0^2 =$ constant, the geometry being given by $\alpha^2 \sin\phi(s)$. Setting $\beta = a_1 sf$, $\psi = a_2 sg$, with suitable constants a_i, observing $|\beta \sin\phi| \ll 1$, and neglecting the small terms $A_i(s)$, (6.1) transforms into (2.1), for $x = s/s_1$.

We now consider two solutions (β_1, ψ_1) and (β_2, ψ_2) of (6.1) and appropriate boundary conditions in terms of β and ψ, which will be discussed later. Setting $v = \beta_1 - \beta_2$, $w = \psi_1 - \psi_2$, multiplying (6.1) by r/α and subtracting, we find

$$D\left(\frac{r}{\alpha}v'\right)' = D\frac{r}{\alpha}\left(\frac{r'}{r}\right)^2 v - \alpha w \sin\phi + \alpha r P_v v \sin\phi + \alpha(\beta_1\psi_1 - \beta_2\psi_2)\cos\phi, \tag{6.3}$$

$$\frac{1}{C}\left(\frac{r}{\alpha}w'\right)' = \frac{r}{C\alpha}\left(\frac{r'}{r}\right)^2 w + \alpha v \sin\phi - \frac{1}{2}\alpha(\beta_1^2 - \beta_2^2)\cos\phi. \tag{6.4}$$

Multiplying (6.3) by v and (6.4) by w and integrating as before, we get

$$\begin{aligned} &-\int_{s_0}^{s_1} \frac{r}{\alpha}\left[D(v')^2 + \frac{1}{C}(w')^2\right] ds + \left(D\frac{r}{\alpha}vv' + \frac{r}{C\alpha}ww'\right)\Big|_{s_0}^{s_1} \\ &\quad = \int_{s_0}^{s_1} \frac{r}{\alpha}\left(\frac{r'}{r}\right)^2\left(Dv^2 + \frac{1}{C}w^2\right) ds + \int_{s_0}^{s_1}\left[v^2 r z' P_v + \frac{1}{2}v^2 r'(\psi_1+\psi_2)\right] ds. \end{aligned} \tag{6.5}$$

The term $(\psi_1+\psi_2)v^2/2$ is obtained exactly as in § 3 due to the same structure of the nonlinearity in (2.1) and (6.1). The integral on the left and the first integral on the right-hand side of the identity (6.5) are nonnegative since $r/\alpha \geqq 0$. Note that in the remaining integral we may have $r' < 0$ and/or $z' < 0$, depending on the given shell geometry. For unrestricted rotations, the stress resultants N_s, N_r, and P_v are related by [11]

$$N_s = N_r \cos(\phi+\beta) + P_v \sin(\phi+\beta).$$

In the approximation represented by (6.1), β is small in the sense that terms of higher than second degree are ignored. Hence we have, in small rotation theory,

$$\begin{aligned} \alpha r N_s &= \alpha r(N_r \cos\phi + P_v \sin\phi) + \alpha r\beta(-N_r \sin\phi + P_v \cos\phi) \\ &= r'\psi + rz'P_v + [-z'\phi\beta + rr'P_v\beta]. \end{aligned} \tag{6.6}$$

Insofar as cubic terms have been neglected in (6.1), there should be no quartic terms in the integrals of (6.5). Hence it is consistent within small finite displacement theory to approximate the second integral on the right-hand side by

$$\int_{s_0}^{s_1} v^2 \alpha r \frac{1}{2}(N_{s1}+N_{s2})\,ds, \tag{6.7}$$

neglecting quartic terms such as $v^2\beta_i\psi_i$ arising from the bracketed term in (6.6). A reasonable concept of tensile solution is stated in the following.

DEFINITION 2. A solution $(\beta(s), \psi(s))$ of (6.1) is called *tensile* if the meridional stress resultant N_s is nonnegative.

The relation (6.6) shows that for shallow shells this concept of a tensile solution is the same as the one introduced in Definition 1, § 3. As the integral (6.7) is nonnegative for tensile solutions, the proof of the uniqueness of solutions (β, ψ), $N_s \geqq 0$, is complete for all boundary conditions that make the term $(\cdots)\big|_{s_0}^{s_1}$ on the left-hand side of (6.5) nonpositive.

Suppose first that β and ψ satisfy the same type of linear boundary conditions as f and g, respectively. Setting $x = s/s_1$ and $\varepsilon = s_0/s_1$, they are given in the form (2.2)-(2.6) with f, g replaced by β, ψ. The BVP's for shells closed at the apex are again denoted as Problem I. In this case we have the symmetry condition $\beta(0)=0$, and $\psi(0)=0$ due to $r(0)=0$, $N_r(0)$ finite. As in § 3, we conclude that the boundary term in (6.5) reduces to

$$\frac{r}{\alpha}\left[Dv(s_1)v'(s_1)+\frac{1}{C}w(s_1)w'(s_1)\right] \leqq 0.$$

Clearly the method of weighted differences given in § 4 can also be applied to the present problem in order to transform the boundary terms in (6.5) such that they become nonpositive for all 16 cases of BVP's for open shells, denoted as Problem II. We pause to summarize these conclusions and state the main result of this section.

THEOREM 6.1. *Tensile solutions* $(\beta(s), \psi(s))$ *in the sense of Definition* 2 *of both Problem* I *and Problem* II *for nonshallow shells are unique if the boundary conditions for* β *and* ψ *at both edges of the shell are of the form*

$$a_1\beta' + a_2\beta = c_1, \qquad a_3\psi' + a_4\psi = c_2 \tag{6.8}$$

where the constants a_i *satisfy* $a_1^2+a_2^2>0$ *and* $a_3^2+a_4^2>0$.

The restrictions notwithstanding, this theorem covers some physically significant BVP's. When β and N_r are prescribed at the boundary, (6.8) is satisfied with $a_1 = a_3 = 0$, which corresponds to the case denoted by $(c, s; C, S)$ in § 4. Furthermore, it will be seen from relations (6.9)-(6.11) that boundary conditions where the meridional bending moment M_s and/or the radial displacement u are prescribed at one or at both edges are linear if $\nu = 0$. Hence, these cases are fully covered by Theorem 6.1.

Next we consider conditions where M_s and/or u is prescribed and $\nu \neq 0$. The meridional bending moment consistent with (6.1) is given by [11]

$$M_s = D\left[\frac{\beta'}{\alpha}+\frac{\nu}{r}\left(\beta\cos\phi+\frac{1}{2}\beta^2\sin\phi\right)\right], \tag{6.9}$$

which leads to a boundary term

$$vv' = -\frac{1}{r}\nu v^2\left[r'+\frac{1}{2}z'(\beta_1+\beta_2)\right]. \tag{6.10}$$

Although our model implies that $z'(\beta_1+\beta_2)$ is in general small compared to r', so long as $r'>0$ and $\cos\phi>(1/2)\sin\phi$, the sign of the term within the brackets of (6.10) cannot be determined a priori. Hence, Theorem 6.1 is rigorously applicable to BVP's with prescribed edge moment only if $z'=0$ and $r'\geqq 0$ at that edge.

We run into a similar difficulty if the displacement is prescribed at the edge. For instance, the radial displacement is given by

$$u=\frac{r}{Eh}(N_\theta-\nu N_s), \qquad N_\theta=\frac{1}{\alpha}\psi'+rp_r. \tag{6.11}$$

This leads to boundary terms $(\beta_1\psi_1-\beta_2\psi_2)z'$ and $rP_v vw$ in (6.5), whose signs cannot be determined except in the special case $z'=P_v=0$ at the edge. The boundary condition $N_s=S$ is also nonlinear and cannot be treated by the present method. Hence, apart from the special cases just mentioned, it remains an open problem to extend Theorem 6.1 to a general class of BVP's, with nonlinear boundary conditions not excluded.

7. Concluding remarks. We have considered axisymmetric deformations under the assumption that the meridional membrane stress is tensile. In the cases we have studied here, we have shown that there cannot be two different tensile solutions. The physical interpretation is that there cannot be symmetric snap buckling, characterized by the occurrence of a limit point, say $\gamma=\gamma_c$, along a load deflection path, unless the radial membrane stress is compressive in the unbuckled or buckled state at $\gamma=\gamma_c$, at least in some part of the shell. It is interesting to note that it has been a general experience in shell design that the bending stresses are unimportant for stability considerations, even when they are not confined to narrow edge layers, as in the case of shallow shells. This observation is confirmed by the results of this paper.

In problems involving finite deformations, tensile membrane stresses cannot, in general, be predicted a priori under a given load. An exception is the circular and the annular plate, where uniqueness of axisymmetric solutions can be proved for arbitrary load without assuming that the solution is tensile (Theorems 3.2 and 5.1). This result can be expected to hold also for very shallow shells, but we have not attempted to prove it. However, a priori uniqueness of solutions can be proved for small loads, provided the shell is sufficiently shallow. This was done by different methods in [18] for arbitrary shallow shells; the results can be adapted to axisymmetric deformations of shells of revolution.

Finally, we remark that it should be possible to include large rotations in the analysis. In particular, we expect uniqueness of axisymmetric solutions for problems of circular and annular plates under arbitrary vertical load. The finite rotation equations for plates [12] are stated here, to indicate that this extension is apparently not straightforward

$$r^2\beta''+r\beta'-\sin\beta\cos\beta=\frac{r}{D}(\psi\sin\beta-P_v\cos\beta),$$

$$r^2\psi''+r\psi'-\psi=\frac{r}{A}(\cos\beta-1)$$

with $r=s$ and primes denoting differentiation with respect to r. In addition to the different structure of the nonlinearity, we have nonlinear boundary conditions in most cases of physical interest, except when $\nu=0$. The above equations can be reduced to a single second order equation in the limit case of circular and annular membranes ($D=0$). Grabmüller and Pirner [21] have recently succeeded in proving the uniqueness of positive solutions for these problems under all physically relevant (linear and

nonlinear) boundary conditions with ν in the range $-1 < \nu \leqq \frac{1}{2}$. For circular membranes with linear boundary conditions, this result was obtained in [20].

Acknowledgments. The authors wishes to thank L. A. Mysak for his warm hospitality during the author's stay in Vancouver.

REFERENCES

[1] L. BAUER, H. B. KELLER, AND E. L. REISS, *Axisymmetric buckling of hollow spheres and hemispheres*, Comm. Pure Appl. Math., 23 (1970), pp. 529-568.

[2] R. W. DICKEY, *Nonlinear bending of circular plates*, SIAM J. Appl. Math. 30 (1976), pp. 1-9.

[3] H. GRABMÜLLER AND E. NOVAK, *Nonlinear boundary value problems for the annular membrane: a note on uniqueness of positive solutions*, J. Elasticity, 17 (1987), pp. 279-284.

[4] ———, *Nonlinear boundary value problems for the annular membrane: new results on existence of positive solutions*, Math. Methods Appl. Sci., (1987), to appear.

[5] H. GRABMÜLLER AND H. J. WEINITSCHKE, *Finite displacements of annular elastic membranes*, J. Elasticity, 16 (1986), pp. 135-147.

[6] G. H. KNIGHTLY AND D. SATHER, *Existence and stability of axisymmetric buckled states of spherical shells*, Arch. Rational Mech. Anal., 63 (1977), pp. 305-319.

[7] W. J. KOITER, *The intrinsic equations of shell theory with some applications*, in Mechanics Today 5 (E. Reissner Anniversary Volume), S. Nemat-Nasser, ed., Pergamon Press, Oxford, New York, 1980, pp. 139-154.

[8] C. G. LANGE AND G. A. KRIEGSMANN, *The axisymmetric branching behavior of complete spherical shells*, Quart. Appl. Math., 39 (1981), pp. 145-178.

[9] A. LIBAI AND J. G. SIMMONDS, *Nonlinear elastic shell theory*, Adv. in Appl. Mech., 23 (1983), pp. 271-371.

[10] E. L. REISS, *A uniqueness theorem for the nonlinear axisymmetric bending of circular plates*, AIAA J., 1 (1963), pp. 2650-2652.

[11] E. REISSNER, *On axisymmetric deformations of thin shells of revolution*, Proc. Sympos. Appl. Math., 3 (1950), pp. 27-52.

[12] ———, *On the equations for finite symmetrical deflections of thin shells of revolution*, in Progress in Appl. Mech. (Prager Anniversary Volume), Macmillan, New York, 1963, pp. 171-178.

[13] E. SCHWERIN, *Über Spannungen und Formänderungen kreisringförmiger Membranen*, Z. Techn. Phys., 12 (1929), pp. 651-659.

[14] L. S. SRUBSHCHIK, *On solvability of nonlinear equations of Reissner type for nonshallow symmetrically loaded shells of revolution*, J. Appl. Math. Mech. (PMM) 32 (1968), pp. 322-326.

[15] N. WAGNER, *Existence theorem for a nonlinear boundary value problem in ordinary differential equations*, Contrib. Differential Equations, 3 (1965), pp. 325-336.

[16] H. J. WEINITSCHKE, *On the stability problem for shallow spherical shells*, J. Math. and Phys., 38 (1959), pp. 209-231.

[17] ———, *On asymmetric buckling of shallow spherical shells*, J. Math. and Phys., 44 (1965), pp. 141-163.

[18] ———, *Some mathematical problems in the non-linear theory of elastic membranes, plates and shells*, in Trends in Applications of Pure Mathematics to Mechanics, G. Fichera, ed., Pitman, London, 1976, pp. 409-424.

[19] ———, *On axisymmetric deformations of nonlinear elastic membranes*, in Mechanics Today 5 (E. Reissner Anniversary Volume), S. Nemat-Nasser, ed., Pergamon Press, Oxford, New York, 1980, pp. 523-542.

[20] ———, *On finite displacements of circular elastic membranes*, Math. Methods Appl. Sci., 9 (1987), pp. 76-98.

[21] H. GRABMÜLLER AND R. PIRNER, *Positive solutions of annular elastic membrane problems with finite rotations*, Stud. Appl. Math., to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

007

A MATHEMATICAL MODEL OF CONTRACTING MUSCLE WITH VISCOELASTIC ELEMENTS*

V. COMINCIOLI† AND A. TORELLI†

Abstract. A three-element model of contracting muscle is studied. This model incorporates a contractile element based on a two-state cross-bridge mechanism and two viscoelastic elements placed in series and in parallel to the contractile element. The existence, uniqueness, and asymptotic behavior of the mathematical solution is proved and the numerical approach is discussed.

Key words. cross-bridge mechanism, muscle contraction, nonlinear nonlocal partial differential equations

AMS(MOS) subject classifications. 92A09, 35F25

1. Introduction. According to the classic view of Hill [15] the muscle's mechanical properties can be separated into three elements: an active force generating contractile element (CE) representing the processes in response to a stimulation and two passive elastic elements: a series elastic element (SE) in series with the (CE), which represents the structures on which (CE) exerts its force during contraction (tendons in skeletal muscles and inactive or less active regions in cardiac muscle) and a parallel elastic component (PE), which determines the mechanical behavior of muscle at rest (sarcolemma of muscle cells and extracellular structures).

On this rheological framework many mathematical models have been introduced that differentiate for the different description of (CE) element. We quote some recent papers in our research group from which more convenient literature can also be found [5], [7], [6], [10], [11], [12], [4].

In these models the (PE) and (SE) elements are described as *elastic* springs following usually an exponential law. There is however a wealth of evidence that such elements exhibit a viscoelastic behavior (see, e.g., [14] and related literature, [2], [18], [3], [19]).

This requires an extension of the traditional Hill model by replacing the series and parallel elastic elements with *viscoelastic* elements. Along this direction we find the model proposed by Glantz [14] in which the (CE) element is described macroscopically on the basis of a *force-velocity curve.*

Our aim in this paper is to introduce in the Glantz's model a more structural description of the (CE) element by making use of the *sliding filament theory.*

According to this theory the generation of muscular force results from interactions between the myosin and actin filaments. Under the influence of the intracellular calcium concentration $[Ca^{2+}]$, which in turn depends on the time course of the action potential, we have the formation of links (cross-bridges) which act like springs.

From the original model proposed by A. F. Huxley in 1957 [16] a number of models have been proposed to account for new experimental findings (for an overview, also see [7], [21], [20], [24], [25]).

However, since in the present paper we want mainly to explore the *mathematical* implications of the replying elastic elements with nonlinear viscoelastic elements, we

* Received by the editors March 3, 1986; accepted for publication (in revised form) March 24, 1987. This work was supported by Ministero Pubblica Istruzione (fondi per la ricerca scientifica) and by Istituto Analisi Numerica del Consiglio Nazionale delle Ricerche, Pavia, Italy.

† Dipartimento di Matematica, Universitá di Pavia, Strada Nuova 65, 27100 Pavia, Italy.

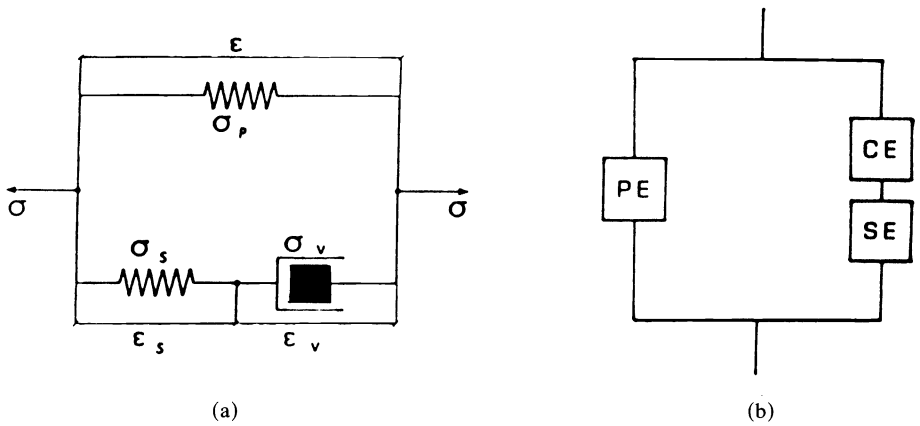
shall consider, for simplicity, the Huxley original model. For the same reason we shall study only the *isometric* contraction, corresponding to the experimental situation in which the muscle's length is kept constant and the force generated is observed. In this case the force of (PE) is constant and we have only to consider the contribution of the force of (SE). In [8], however, we report some numerical results corresponding to more general situations: *isotonic* (the force is assigned and muscle length is computed) and *isometric-isotonic.*

The present paper represents a *first* contribution in the validation process of the proposed model; indeed we prove that it is *mathematically* consistent, that is *there exists a unique solution* for admissible physical data, which agrees qualitatively with observed phenomena. We are now *identifying* the model parameters using experimental data collected in the muscle physiology laboratory at the Institute of Human Physiology, Pavia, Italy. The comparison with the results obtained by means of models containing only elastic passive elements ([5], [7], [6], [20], [23]) will be useful to estimate in quantitative ways the contribution of the viscosity in passive elements.

2. The model. The muscle, supposed homogeneous, is represented (Fig. 1(b)) by a three-element model consisting of a contractile element (CE) and two passive viscoelastic elements (SE), (PE), like those represented in Fig. 1(a) (Kelvin model, slightly different from that used in [14]).

Each of these elements is described by the following equation:

$$
\begin{aligned}
&\sigma_v = \eta \varepsilon_v' (\text{dashpot}), \qquad \varepsilon' = \frac{d\varepsilon}{dt}, \\
&\sigma_p = a(\exp(b(\varepsilon - \varepsilon_0)) - 1), \qquad \sigma_s = a(\exp(b(\varepsilon_s - \varepsilon_{s0})) - 1), \\
&\sigma_v = \sigma_s, \qquad \sigma_p + \sigma_v = \sigma, \\
&\varepsilon_v + \varepsilon_s = \varepsilon
\end{aligned}
\tag{2.1}
$$



(a) (b)

FIG. 1. (a) *Arrangement of pure elastic and viscous elements.* (b) *Classical three-element model consisting of a contractile element and two viscoelastic passive elements.*

where σ = force, ε = length, ε_0, ε_{s0} rest length. For simplicity we have supposed that the constants a, b are the same for both forces σ_p, σ_s.

From (2.1), eliminating the "internal" variables σ_v, ε_v, ε_s, σ_s, we have the following *constitutive* equation for a viscous element:

$$\sigma' = \varepsilon'(2ab + b\sigma) - \frac{b}{\eta}(\sigma - \sigma_p)(a + \sigma - \sigma_p). \tag{2.2}$$

We note that for $\eta = 0$ and for $\eta \to +\infty$ we have again two (different) only elastic situations.

We consider now the contractile element (CE). As already said it is described, for simplicity, on the basis of Huxley's theory [16] (see also for more details [5]). This element is identified with the half-sarcomere, which is the repeating unit of muscle structure and consists of an array of the thick (myosin) and the thin (actin) filaments. The links (cross-bridges) between these filaments are characterized by the distance x between the equilibrium position of the myosin head and reactive site (Fig. 2).

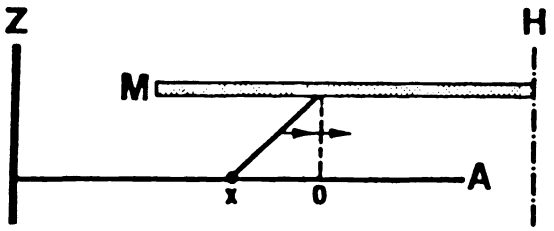


FIG. 2. *Schematic organization of a sarcomere*: A, *actin*; M, *myosin*; Z–*line*.

Let $n(x, t)$ denote the relative cross-bridge density at time t (fraction of the attached cross-bridges per unit of cross-bridge length in one half-sarcomere).

Supposing the cross-bridges to behave as linear elastic bonds with stiffness k and $n(x, t) = 0$ for x great enough, the force developed at time t by the half-sarcomere is given by

$$FCE(t) = k \int_{-\infty}^{+\infty} n(\xi, t)\xi \, d\xi. \tag{2.3}$$

The dynamics of the cross-bridges population $n(x, t)$ results from the balance of the formation and breakage, that is,

$$\frac{dn(x, t)}{dt} = f(x)\gamma(t)(1 - n(x, t)) - g(x)n(x, t) \tag{2.4}$$

where $\gamma(t)$ is the activation function, $f(x)$, $g(x)$ are the attachment rate functions and d/dt denotes the material derivative, i.e., the derivative with respect to a frame moving with the cross-bridges distribution.

Then we have

$$\frac{d}{dt} = \frac{\partial}{\partial t} + v\frac{\partial}{\partial x}, \qquad v = \frac{dx}{dt};$$

v is the half-sarcomere *velocity shortening*, that is,

$$v = \frac{dLCE}{dt}. \tag{2.5}$$

Since the sarcomere shortening follows the formation of cross-bridges, we have that v is a function of n. We want now to specify this dependence. From the rheological model we have at any time t verified these conditions:

$$\begin{aligned} &LCE(t)+LSE(t)=L(t),\\ &FCE(t)=FSE(t),\\ &P(t)=FCE(t)+FPE(t) \end{aligned} \tag{2.6}$$

where LCE, LSE, L represent the lengths of each element and FCE, FPE, P the forces.

In the following we shall consider the *isometric* situation: $L(t)=\text{const}$ and $P(t)$ is the output model. We have

$$\frac{dLCE}{dt}=-\frac{dLSE}{dt}. \tag{2.7}$$

From (2.1), noting that $\sigma=FSE$, $\varepsilon=LSE$, we have Problem 2.1.

PROBLEM 2.1. *$n(x,t)$ is the solution for $t>0$ and $x\in\mathbf{R}$ of the following equation*:

$$\frac{\partial n}{\partial t}+v\frac{\partial n}{\partial x}=f(x)\gamma(t)(1-n)-g(x)n \tag{2.8}$$

where $v(t)=-\varepsilon'(t)$ with $\varepsilon(t)$ solution of the differential equation

$$\sigma'=\varepsilon'(2ab+b\sigma)-\frac{b}{\eta}(\sigma-\sigma_p)(a+\sigma-\sigma_p) \tag{2.9}$$

where

$$\sigma(t)=k\int_{-\infty}^{+\infty} n(\xi,t)\xi\,d\xi. \tag{2.10}$$

Assuming at $t=0$ a resting situation, we have the initial conditions

$$n(x,0)=0, \qquad \varepsilon(0)=\varepsilon_0.$$

We shall prove in the following that Problem 2.1 has a unique solution under general assumptions on data $\gamma(t)$, $f(x)$, $g(x)$ and we shall study the behavior of the solution for $\eta\to 0$ and $\eta\to+\infty$.

To simplify the notation, but without loss of generality, we set the constants a, b, k equal to one for the following. Then we indicate by $u(x,t)$ the unknown function $n(x,t)$ and by $z(t)$ the function: $-(\varepsilon(t)-\varepsilon_0)$, that is the variation of the LCE length.

3. Mathematical formulation of the problem. We assume that

$$f, g\in C^1(\mathbf{R}), \qquad \gamma\in C^1([0,+\infty[), \tag{3.1}$$

$$f(x), g(x)\geqq 0, \qquad x\in\mathbf{R}, \tag{3.2}$$

$$\gamma(t)\geqq 0, \qquad t\in[0,+\infty[, \tag{3.3}$$

(3.4) The support of f is a compact set of $\mathbf{R}$.

Let also T be a fixed nonnegative number. According to § 2, we introduce the following.

PROBLEM 3.1. *Given $\eta\in[0,+\infty]$, we look for a couple $\{u,z\}$ verifying*

$$u\in C^1(\mathbf{R}\times[0,T]), \qquad z\in C^1([0,T]), \tag{3.5}$$

$$u(x,0)=0, \quad x\in\mathbf{R}; \qquad z(0)=0, \tag{3.6}$$

(3.7) *for all* $t \in [0, T]$, *the support of* $u(\cdot, t)$ *is a compact set of* $\mathbf{R}$,

$$u_t + z' u_x = \gamma f(1-u) - gu; \tag{3.8}$$

moreover, if we put

$$\sigma(t) = \int_{\mathbf{R}} xu(x, t)\, dx, \tag{3.9}$$

the following relations must also be fulfilled:

(i) *If* $\eta \in]0, +\infty[$, *then*

$$\sigma + 2 - \exp(-z) > 0, \tag{3.10}$$

$$\frac{[\sigma+1-\exp(-z)][\sigma+2-\exp(-z)]}{\sigma+2} = -\eta\left(z' + \frac{\sigma'}{\sigma+2}\right). \tag{3.11}$$

(ii) *If* $\eta = 0$, *then*

$$\sigma + 1 - \exp(-z) = 0. \tag{3.12}$$

(iii) *If* $\eta = +\infty$, *then*

$$\sigma + 2 - 2\exp(-z) = 0. \tag{3.13}$$

Remark 3.1. To simplify the notation we have omitted the dependence on η.

To study Problem 3.1 the idea is to introduce an *equivalent* problem (see next Problem 4.2) in terms of an *integral equation* for the function z; then this equation is studied by a fixed point argument.

4. Preliminary considerations. (a) If $\{u, z\}$ is a solution of Problem 3.1, we put

$$p(t) = [\sigma(t) + 2 - \exp(-z)]^{-1}, \tag{4.1}$$

and we have the following:

(i) If $\eta = 0$, then

$$p(t) \equiv 1. \tag{4.2}$$

(ii) If $\eta = +\infty$, then

$$p(t) = \exp(z(t)). \tag{4.3}$$

(iii) If $\eta \in]0, +\infty[$, then

$$p(0) = 1, \tag{4.4}$$

$$p(t) > 0, \qquad t \in [0, T], \tag{4.5}$$

$$p' = pz' + \frac{1}{\eta}(1-p). \tag{4.6}$$

(b) We now introduce the following nonlinear operator:

$$S: [0, +\infty] \times C^0([0, T]) \to C^0([0, T]) \tag{4.7}$$

for $r \in C^0([0, T])$ defined as follows:

$$S[0, r] \equiv 1, \tag{4.8}$$

$$S[+\infty, r] = \exp(r), \tag{4.9}$$

and, if $\eta \in]0, +\infty[$,

$$S[\eta, r](t) = \exp\left(r(t) - \frac{1}{\eta}t\right) + \frac{1}{\eta}\int_0^t \exp\left(r(t) - r(s) + \frac{s-t}{\eta}\right) ds. \tag{4.10}$$

It is easy to prove.

PROPOSITION 4.1. *For every $\eta \in [0, +\infty]$ and $r \in C^0([0, T])$ we have that*

$$S[\eta, r](t) > 0 \quad \forall t \in [0, T]. \tag{4.11}$$

PROPOSITION 4.2. *For every $\eta \in]0, +\infty[$ and $r \in C^1([0, T])$, $r(0) = 0$, the function $p = S[\eta, r]$ is the unique solution of problem* (4.4) *and* (4.6), *with z replaced by r.*

(c) We can now give a new formulation of Problem 3.1.

PROBLEM 4.1. *Given $\eta \in [0, +\infty]$, we look for a couple $\{u, z\}$ verifying* (3.5)-(3.8) *and*

$$S[\eta, z] = \left[2 - \exp(-z) + \int_{\mathbf{R}} xu(x, t)\, dx\right]^{-1}. \tag{4.12}$$

Problems 3.1 and 4.1 are equivalent as stated by the following proposition (the proof of which is immediate).

PROPOSITION 4.3. *Let $\eta \in [0, +\infty]$. We have that $\{u, z\}$ is a solution of Problem* 3.1 *if and only if it is a solution of Problem* 4.1.

(d) Now let

$$H(x, t) = \gamma(t)f(x) + g(x). \tag{4.13}$$

Following [5], we introduce the following operator (where $r \in C^0([0, T])$):

$$\begin{aligned}(Ur)(x, t) = &\int_0^t \gamma(s) f(r(s) - r(t) + x) \\ &\cdot \exp\left[-\int_s^t H(r(\tau) - r(t) + x, \tau)\, d\tau\right] ds.\end{aligned} \tag{4.14}$$

Given $r \in C^0([0, T])$, we put

$$[r](t) = \max\{|r(s) - r(0)|, s \in [0, t]\}. \tag{4.15}$$

If $v \in C^0(\mathbf{R} \times [0, T])$, we put

$$\operatorname{supp}_t(v) = \overline{\{x \in \mathbf{R}: v(x, t) \neq 0\}}, \tag{4.16}$$

that is, $\operatorname{supp}_t(v)$ is the support of v, with respect to x, at the time t.

Thanks to (3.4), we can find $N_-, N_+ \in \mathbf{R}$ such that

$$\operatorname{supp}(f) \subset [-N_-, N_+], \qquad N_-, N_+ > 0. \tag{4.17}$$

The following result can be proved by the method of [5].

PROPOSITION 4.4. (i) *Let $r \in C^0([0, T])$. Then we have*

$$0 \leqq (Ur)(x, t) \leqq 1, \qquad (x, t) \in \mathbf{R} \times [0, T], \tag{4.18}$$

$$\operatorname{supp}_t(Ur) \subset [-N_- - 2[r](t), N_+ + 2[r](t)]. \tag{4.19}$$

(ii) *If $r \in C^1([0, T])$, then $v = Ur$ is the unique solution of the problem*

$$v_t + r'v_x = \gamma f - Hv, \qquad v(x, 0) = 0. \tag{4.20}$$

(e) If $r \in C^0([0, T])$, we define

$$(Ar)(t) = \int_{\mathbf{R}} x[\gamma f - H(Ur)]\, dx. \tag{4.21}$$

The operator A has a meaning by (3.4) and by Proposition 4.4. Also let

$$F: [0, +\infty] \times C^0([0, T]) \to C^0([0, T]), \tag{4.22}$$

defined by

$$F[0, r] = \frac{-Ar}{\exp(-r) + \int_{\mathbf{R}} (Ur)(x, t)\, dx}, \tag{4.23}$$

$$F[+\infty, r] = \frac{-Ar}{2\exp(-r) + \int_{\mathbf{R}} Ur(x, t)\, dx} \tag{4.24}$$

and for every $\eta \in]0, +\infty[$

$$F[\eta, r] = -\frac{Ar + (1/\eta)((1 - S[\eta, r])/(S[\eta, r])^2)}{(S[\eta, r])^{-1} + \exp(-r) + \int_{\mathbf{R}} (Ur)(x, t)\, dx}. \tag{4.25}$$

The operator F has a meaning by Propositions 4.1 and 4.4. For every $\eta \in [0, +\infty]$ and $r \in C^0([0, T])$, we define

$$W[\eta, r](t) = \int_0^t F[\eta, r](s)\, ds. \tag{4.26}$$

We can now state the following problem.

PROBLEM 4.2. *Given $\eta \in [0, +\infty]$, we look for a function z verifying*

$$z \in C^1([0, T]), \tag{4.27}$$

$$z = W[\eta, z]. \tag{4.28}$$

Problems 4.1 and 4.2 are equivalent as stated by Proposition 4.5.

PROPOSITION 4.5. *Let $\eta \in [0, +\infty]$. Then*

(i) *If $\{u, z\}$ is a solution of Problem* 4.1, *then z is a solution of Problem* 4.2.

(ii) *If z is a solution of Problem* 4.2, *then the couple $\{Uz, z\}$ is a solution of Problem* 4.1.

Proof. (i) Let $\{u, z\}$ be a solution of Problem 4.1. By (3.8) and by Proposition 4.4, it follows that $u = Uz$. Multiplying (3.8) by x and integrating on $\mathbf{R}$, we have

$$\sigma' - z' \int_{\mathbf{R}} (Uz)(x, t)\, dx = Az, \tag{4.29}$$

where σ is defined in (3.9) and Az is defined in (4.21). Now put

$$p = [\sigma + 2 - \exp(-z)]^{-1}. \tag{4.30}$$

By (4.12), we obtain

$$p = S[\eta, z]; \tag{4.31}$$

hence (by Proposition 4.2),

$$p' = pz' + \frac{1}{\eta}(1 - p), \qquad \eta \in]0, +\infty[. \tag{4.32}$$

Recalling (4.29) and (4.30), we have

$$z'\left[\exp(-z)+\int_{\mathbf{R}}(Uz)(x,t)\,dx\right]=-\frac{p'}{p^2}-Az. \tag{4.33}$$

Using (4.8) and (4.9), we immediately obtain ($\eta=0,+\infty$), respectively,

$$z'=F[0,z],\qquad z'=F[+\infty,z]. \tag{4.34}$$

Let now $\eta\in\,]0,+\infty[$. It follows from (4.25), (4.32), and (4.33) that

$$z'=F[\eta,z],\qquad \eta\in\,]0,+\infty[. \tag{4.35}$$

Relations (4.34) and (4.35) imply (4.28).

(ii) Let z be a solution of Problem 4.2. Setting $u=Uz$, we must prove that $\{u,z\}$ verifies (3.5)-(3.8) and (4.12). The relations (4.14) and (4.26) imply (3.5) and (3.6). The relations (3.7) and (3.8) are a consequence of Proposition 4.4. It remains to prove (4.12). Assume $\eta\in\,]0,+\infty[$. Let

$$p=S[\eta,z]. \tag{4.36}$$

By Proposition 4.2 and (4.28), we have

$$z'\left[\exp(-z)+\int_{\mathbf{R}}u(x,t)\,dx\right]=-Az-\frac{p'}{p^2}. \tag{4.37}$$

By (3.8) (already proved), after multiplication by x and integration on $\mathbf{R}$, we have

$$\left(\int_{\mathbf{R}}xu(x,t)\,dx\right)'-z'\int_{\mathbf{R}}u(x,t)\,dx=Az. \tag{4.38}$$

Adding (4.37) and (4.38), recalling the initial data given by (3.6) and using Proposition 4.2, it follows that

$$-\exp(-z)+\int_{\mathbf{R}}xu(x,t)\,dx=\frac{1}{p}-2, \tag{4.39}$$

that is (4.12), in the special case $\eta\in\,]0,+\infty[$. If $\eta\in\{0,+\infty\}$ the proof is similar.

5. Existence and uniqueness of the solution. (a) Now set $r\in C^0([0,T])$,

$$\bar{r}(t)=\max\{r(s),\,s\in[0,t]\}, \tag{5.1}$$

$$\underline{r}(t)=\min\{r(s),\,s\in[0,t]\}, \tag{5.2}$$

$$E_r^+=\{t\in[0,T]:r(t)=\bar{r}(t)\}, \tag{5.3}$$

$$E_r^-=\{t\in[0,T]:r(t)=\underline{r}(t)\}. \tag{5.4}$$

LEMMA 5.1. *If* $\eta\in\,]0,+\infty[$ *and* $r\in C^0([0,T])$, *it follows that*

$$\begin{aligned}\exp(-\bar{r}(t))\left[1-\exp\left(\frac{-t}{\eta}\right)\right]+\exp\left(\frac{-t}{\eta}\right)&\leqq\frac{S[\eta,r](t)}{\exp(r(t))}\\&\leqq\exp(-\underline{r}(t))\left[1-\exp\left(\frac{-t}{\eta}\right)\right]+\exp\left(\frac{-t}{\eta}\right).\end{aligned} \tag{5.5}$$

Proof. Recalling (4.10), we have

$$S[\eta, r](t) \leqq \exp\left(r(t)-\frac{t}{\eta}\right)\left[1+\frac{1}{\eta}\exp\left(-\underline{r}(t)\right)\int_0^t \exp\frac{s}{\eta}\,ds\right]$$

$$= \exp\left(r(t)-\frac{t}{\eta}\right)\left[1+\exp\left(-\underline{r}(t)\right)\left(\exp\frac{t}{\eta}-1\right)\right].$$

The opposite inequality can be proved in a similar way.

COROLLARY 5.1. *Let* $\eta \in [0, +\infty]$ *and* $r \in C^0([0, T])$, *with* $r(0) = 0$. *Then we have*

(5.6) *if* $t \in E_r^+$, *then* $S[\eta, r](t) \geqq 1$,

(5.7) *if* $t \in E_r^-$, *then* $S[\eta, r](t) \leqq 1$.

Proof. The cases $\eta = 0$ and $\eta = +\infty$ are obvious. Let now $\eta \in]0, +\infty[$ and $t \in E_r^+$. We have that $r(t) = \bar{r}(t)$. Hence (by Lemma 5.1),

$$S[\eta, r](t) \geqq 1 - \exp\left(\frac{-t}{\eta}\right) + \exp\left(\bar{r}(t)-\frac{t}{\eta}\right).$$

Recalling that $r(0) = 1$, it follows that $\bar{r}(t) \geqq 0$. We have so obtained (5.6). The proof of (5.7) is similar.

(b) Now put

$$\Gamma(t) = \int_0^t \gamma(s)\,ds. \tag{5.8}$$

We can now prove the following lemma.

LEMMA 5.2. *Let* $\eta \in [0, +\infty]$ *and* z *is a solution of Problem* 4.2. *Then we have* ($t \in [0, T]$)

$$z(t) \geqq -\log\left[1 + N_+\Gamma(t)\int_0^{+\infty} f(x)\,dx\right], \tag{5.9}$$

where N_+ *is defined in* (4.17).

Proof. Let $\bar{t} \in E_z^-$. Put

$$I(\bar{t}) = \int_{-\infty}^{+\infty} x(Uz)(x, \bar{t})\,dx. \tag{5.10}$$

Recalling (3.2), (3.3), (4.13), and (4.14), it follows that

$$I(\bar{t}) \leqq \int_0^{+\infty} x \int_0^{\bar{t}} \gamma(s) f(z(s) - z(\bar{t}) + x)\,ds\,dx.$$

Changing the order of the integration and setting $y = z(s) - z(\bar{t}) + x$, we have

$$I(\bar{t}) \leqq \int_0^{\bar{t}} \gamma(s) \int_{\alpha(s)}^{+\infty} (y - \alpha(s)) f(y)\,dy\,ds,$$

where $\alpha(s) = z(s) - z(\bar{t})$. Since $\bar{t} \in E_z^-$, it follows that $\alpha(s) \geqq 0$ ($s \in [0, \bar{t}]$). Hence,

$$I(\bar{t}) \leqq \int_0^{\bar{t}} \gamma(s) \int_0^{+\infty} y f(y)\,dy\,ds \leqq N_+\Gamma(\bar{t}) \int_0^{+\infty} f(x)\,dx.$$

Recalling now the equivalence between Problems 4.1 and 4.2 (see Proposition 4.5), it follows that

$$(S[\eta, z](\bar{t}))^{-1}+\exp(-z(\bar{t}))-2 \leqq N_+\Gamma(\bar{t}) \int_0^{+\infty} f(y)\, dy.$$

Since $\bar{t} \in E_z^-$, by Corollary 5.1, we have that $S[\eta, z](\bar{t}) \leqq 1$. Hence,

$$\exp(-z(\bar{t})) \leqq 1+N_+\Gamma(\bar{t}) \int_0^{+\infty} f(y)\, dy.$$

This proves Lemma 5.2 in the case $t \in E_z^-$. If $t \in [0, T]$, then there exists $\bar{t} \in E_z^-$, such that

$$0 \leqq \bar{t} \leqq t, \qquad z(\bar{t}) \leqq z(t).$$

It follows that $\Gamma(t) \geqq \Gamma(\bar{t})$, and

$$z(t) \geqq z(\bar{t}) \geqq -\log\left(1+N_+\Gamma(t) \int_0^{+\infty} f(y)\, dy\right),$$

as we needed to show.

(c) Now let

$$K \in \mathbf{N}, \qquad K \geqq N_+\Gamma(T) \int_{-\infty}^0 f(x)\, dx+\frac{1}{1-\exp(-N_-)}, \tag{5.11}$$

where N_+ and N_- are defined in (4.17). Then we have Lemma 5.3.

LEMMA 5.3. *Let $\eta \in [0, +\infty]$. If z is a solution of Problem* 4.2, *then*

$$z(t) \leqq KN_-. \tag{5.12}$$

Proof. By contradiction, we can put $(k = 0, 1, \cdots, K)$

$$t_k = \min\{t \in [0, T]: z(t) = kN_-\}. \tag{5.13}$$

Then we have

$$0 = t_0 < t_1 < \cdots < t_K, \tag{5.14}$$

$$z(t_k) = kN_-, \qquad t_k \in E_z^+. \tag{5.15}$$

Now put

$$I_k = \int_{-\infty}^{+\infty} x(Uz)(x, t_k)\, dx. \tag{5.16}$$

If $k \geqq 1$, it follows that

$$I_k \geqq J_1+J_2, \tag{5.17}$$

where

$$J_1 = \int_{-\infty}^0 x \int_0^{t_{k-1}} \gamma(s) f(z(s)-kN_-+x)\, ds\, dx,$$

$$J_2 = \int_{-\infty}^0 x \int_{t_{k-1}}^{t_k} \gamma(s) f(z(s)-kN_-+x)\, ds\, dx.$$

We have that

$$J_1 = 0. \tag{5.18}$$

Indeed if $s \in [0, t_{k-1}]$, it follows that $z(s) \leqq (k-1)N_-$; hence, if $x \leqq 0$,

$$z(s)-kN_-+x \leqq -N_-.$$

Recalling (4.17), we obtain (5.18). Setting $y = z(s) - kN_- + x$, we have

$$J_2 = \int_{t_{k-1}}^{t_k} \gamma(s) \int_{-N_-}^{\beta(s)} [y - \beta(s)] f(y)\, dy\, ds,$$

where $\beta(s) = z(s) - kN_-$. Since $s \in [t_{k-1}, t_k]$, it follows that $\beta(s) \leqq 0$. Then we have

$$J_2 \geqq \int_{t_{k-1}}^{t_k} \gamma(s) \int_{-N_-}^{0} y f(y)\, dy\, ds;$$

hence

$$J_2 \geqq -N_-[\Gamma(t_k) - \Gamma(t_{k-1})] \int_{-\infty}^{0} f(x)\, dx. \tag{5.19}$$

By (5.16)-(5.19), it follows that

$$\int_{-\infty}^{+\infty} x(Uz)(x, t_k)\, dx \geqq -N_-[\Gamma(t_k) - \Gamma(t_{k-1})] \int_{-\infty}^{0} f(y)\, dy. \tag{5.20}$$

Using the equivalence between Problems 4.1 and 4.2 and recalling Corollary 5.1 (since $t_k \in E_z^+$), we obtain that

$$1 \leqq \exp(-kN_-) + N_-[(\Gamma(t_k) - \Gamma(t_{k-1}))] \int_{-\infty}^{0} f(y)\, dy.$$

Adding with respect to k $(k = 1, \cdots, K)$, we have

$$K \leqq \exp(-N_-) \frac{1 - \exp(-KN_-)}{1 - \exp(-N_-)} + N_-\Gamma(T) \int_{-\infty}^{0} f(y)\, dy,$$

hence

$$K < \frac{1}{1 - \exp(-N_-)} + N_-\Gamma(T) \int_{-\infty}^{0} f(y)\, dy,$$

which contradicts the (5.11).

(d) Notice that the estimates of Lemmas 5.2 and 5.3 are independent of $\eta \in [0, +\infty]$. Recalling also Lemma 5.1 we obtain immediately the following.

LEMMA 5.4. *If z is a solution of Problem* 4.2, *then there exists a constant $c_1 > 0$ (independent of $\eta \in [0, +\infty]$), such that*

$$|z(t)| < c_1, \qquad t \in [0, T], \tag{5.21}$$

$$\exp(-2c_1) \leqq S[\eta, z](t) \leqq \exp(2c_1), \qquad t \in [0, T]. \tag{5.22}$$

(e) It is easy to prove Lemma 5.5.

LEMMA 5.5. *There exists a constant $c_2 > 0$ such that for $r \in C^0([0, T])$, with*

$$|r(t)| \leqq c_1, \qquad t \in [0, T], \tag{5.23}$$

we have that

$$|(Ar)(t)| \leqq c_2, \qquad t \in [0, T], \tag{5.24}$$

where Ar is defined in (4.21) *and c_1 is the constant appearing in Lemma* 5.4.

Proof. It is a consequence of (3.1)-(3.4) and of Proposition 4.4.

(f) Looking at the formulation of Problem 4.2 and recalling Proposition 4.4 and Lemma 5.1, it is easy to find $c_3(\eta) > 0$ such that

$$\left| \frac{d}{dt} W[\eta, r](t) \right| \leqq c_3(\eta), \tag{5.25}$$

for all $r \in C^0([0, T])$ verifying

$$|r(t)| \leqq c_1, \qquad t \in [0, T]. \tag{5.26}$$

For more details, see [5]. Always following [5] (see also [7]), we can now apply the theorem of Browder [1] or (equivalently) the degree theory to obtain the following existence and uniqueness result.

THEOREM 5.1. *Let* $\eta \in [0, +\infty]$. *Then there exists one and only one solution of Problem* 4.2.

Remark 5.1. Thanks to (5.25) we can find a bound for z' depending on $\eta \in [0, +\infty]$. Having in mind to prove a theorem on the asymptotic behavior of the solution for $\eta \downarrow 0$ and for $\eta \to +\infty$, we need a stronger estimate. We shall do this in the next section.

6. Asymptotic behavior. (a) From now on we specify the dependence on η, writing z_η for a solution of Problem 4.2 with $\eta \in [0, +\infty]$. Also put

$$u_\eta = Uz_\eta, \tag{6.1}$$

$$p_\eta = S[\eta, z_\eta]. \tag{6.2}$$

With the new notation, Problem 4.2 can be written in the following way:

$$z_0' = -\frac{Az_0}{\exp(-z_0) + \int_{\mathbf{R}} u_0(x, t)\, dx}, \tag{6.3}$$

$$z_\infty' = -\frac{Az_\infty}{2\exp(-z_\infty) + \int_{\mathbf{R}} u_\infty(x, t)\, dx}, \tag{6.4}$$

$$z_\eta' = -\frac{Az_\eta + 1/\eta((1-p_\eta)/p_\eta^2)}{p_\eta^{-1} + \exp(-z_\eta) + \int_{\mathbf{R}} u_\eta(x, t)\, dx}, \qquad \eta \in]0, +\infty[\tag{6.5}$$

with the initial condition

$$z_\eta(0) = 0, \qquad \eta \in [0, +\infty]. \tag{6.6}$$

Since p_η is defined by (6.2), it follows that

$$p_\eta' = p_\eta z_\eta' + \frac{1}{\eta}(1 - p_\eta), \qquad p(0) = 1, \qquad \eta \in]0, +\infty[, \tag{6.7}$$

$$p_0 \equiv 1, \qquad p_\infty = \exp(z_\infty). \tag{6.8}$$

Later the following relation will be useful:

$$z_\eta' = -\frac{Az_\eta + p_\eta^{-2} p_\eta'}{\exp(-z_\eta) + \int_{\mathbf{R}} u_\eta(x, t)\, dx}, \qquad \eta \in]0, +\infty[\tag{6.9}$$

obtained by (6.5) and (6.7).

Now let

$$\mu_\eta = \exp(-z_\eta). \tag{6.10}$$

Relation (6.5) can be written as

$$\mu_\eta' = \mu_\eta \frac{Az_\eta + (1/\eta)((1-p_\eta)/p_\eta^2)}{p_\eta^{-1} + \mu_\eta + \int_{\mathbf{R}} u_\eta(x, t)\, dx}, \qquad \eta \in]0, +\infty[. \tag{6.11}$$

Recalling (4.10) and (6.2), we have

$$p_\eta(t) = \frac{1}{\mu_\eta(t)} \exp\left(\frac{-t}{\eta}\right) + \frac{1}{\eta} \int_0^t \left(\frac{\mu_\eta(s)}{\mu_\eta(t)} \exp\frac{s-t}{\eta}\right) ds. \tag{6.12}$$

Integrating by parts, it follows that

$$p_\eta(t)-1=-\int_0^t \frac{\mu'_\eta(s)}{\mu_\eta(t)}\exp\frac{s-t}{\eta}\,ds, \qquad \eta\in\,]0,+\infty[. \tag{6.13}$$

Recalling (6.11), we have

$$\mu'_\eta=\mu_\eta\frac{Az_\eta+(1/\eta)p_\eta^{-2}\mu_\eta^{-1}\int_0^t \mu'_\eta(s)\exp\left((s-t)/\eta\right)ds}{p_\eta^{-1}+\mu_\eta+\int_{\mathbf{R}} u_\eta(x,t)\,dx}. \tag{6.14}$$

(b) We begin the study of the asymptotic behavior of the solution with the following lemma.

LEMMA 6.1. *We have that*

$$\lim_{\eta\downarrow 0}\frac{1}{p_\eta}=1 \quad \textit{weakly in } L^2(]0,T[), \tag{6.15}$$

$$\lim_{\eta\to+\infty} p_\eta\exp(-z_\eta)=1 \quad \textit{uniformly in } [0,T]. \tag{6.16}$$

Proof. If $\eta\in\,]0,+\infty[$, we have that

$$p'_\eta=p_\eta z'_\eta+\frac{1}{\eta}(1-p_\eta). \tag{6.17}$$

Hence

$$\log p_\eta(t)-z_\eta(t)=\frac{1}{\eta}\int_0^t\left(\frac{1}{p_\eta(s)}-1\right)ds. \tag{6.18}$$

Recalling Lemma 5.4, it follows that $(\eta\downarrow 0)$

$$\int_0^t\left(\frac{1}{p_\eta(s)}-1\right)ds\to 0, \qquad t\in[0,T]. \tag{6.19}$$

Hence

$$\frac{1}{p_\eta}\to 1 \quad \text{in } \mathscr{D}'(]0,T[). \tag{6.20}$$

Let now η_n be a sequence such that $\eta_n\downarrow 0$. By Lemma 5.4, the sequence $1/p_{\eta_n}$ is bounded in $L^2(]0,T[)$. Then there exists a subsequence $1/p_{\eta_{n_k}}$ which converges to $l\in L^2(]0,T[)$ (weakly). But, thanks to (6.20), we obtain that $l=1$. Hence

$$\lim_{\eta\downarrow 0}\frac{1}{p_\eta}=1 \quad \text{weakly in } L^2(]0,T[),$$

that is,

$$\lim_{\eta\downarrow 0}\int_0^T\left(\frac{1}{p_\eta}(s)-1\right)\phi(s)\,ds=0 \quad \forall\phi\in L^2(]0,T[).$$

This proves (6.15). Let us now prove (6.16). Recalling Lemmas 5.1 and 5.4, it follows that

$$p_\eta(t)\exp(-z_\eta(t))-1\leqq\left[1-\exp\left(-\frac{t}{\eta}\right)\right][\exp c_1-1];$$

hence

$$p_\eta(t)\exp(-z_\eta(t))-1\leqq\frac{t}{\eta}\,[\exp c_1-1]. \tag{6.21}$$

In a similar way, we have

$$p_\eta(t)\exp(-z_\eta(t))-1\geqq\frac{t}{\eta}[\exp(-c_1)-1]. \tag{6.22}$$

This proves (6.16).

Now we prove an estimate on z'_η, for η not too small.

LEMMA 6.2. *For every $c_3>0$, there exists $c_4>0$ (independent of η), such that*

$$|z'_\eta(t)|\leqq c_4 \quad \forall t\in[0,T] \quad \forall \eta\in\,]c_3,+\infty]. \tag{6.23}$$

Proof. It is an easy consequence of Proposition 4.4 and Lemmas 5.4 and 5.5.

(c) It will be more difficult to prove the estimate on z'_η for η small. We begin by proving Lemma 6.3.

LEMMA 6.3. *There exists a constant $c_5>0$ (independent of $\eta\in[0,+\infty]$) such that*

$$z'_\eta(t)\leqq c_5 \quad \forall t\in[0,T]. \tag{6.24}$$

Proof. The relation (6.24) is obvious for $\eta=0,+\infty$. Now we assume that $\eta\in\,]0,+\infty[$. Thanks to (6.10) and Lemma 5.4, it is sufficient to prove that

$$\mu'_\eta(t)\geqq c_6 \quad \forall t\in[0,T], \tag{6.25}$$

where c_6 is independent of η. Let us fix now $\eta\in\,]0,+\infty[$ and put ($\bar t$ may depend on η)

$$\bar t\in[0,T]:\mu'_\eta(\bar t)\leqq\mu'_\eta(t), \quad t\in[0,T]. \tag{6.26}$$

We can obviously assume that $\mu'_\eta(\bar t)\leqq 0$.

We now distinguish two cases.

Case 1. We assume that $p_\eta(\bar t)\leqq 1$. In this case, using (6.11), we have

$$\mu'_\eta(\bar t)\geqq\mu_\eta(\bar t)\frac{(Az_\eta)(\bar t)}{p_\eta^{-1}(\bar t)+\mu_\eta(\bar t)+\int_{\mathbf{R}}u_\eta(x,\bar t)\,dx}. \tag{6.27}$$

Thanks to Proposition 4.4 and to Lemmas 5.4 and 5.5, this means that there exists c'_6 (independent of η) such that

$$\text{if } p_\eta(\bar t)\leqq 1, \text{ then } \mu'(\bar t)\geqq c'_6. \tag{6.28}$$

Case 2. We assume that $p_\eta(\bar t)>1$. Recalling Proposition 4.4 and the relation (6.14), it follows that there exists a constant C (independent of η) such that

$$\mu'_\eta(\bar t)\geqq C+\frac{1}{\eta}\frac{\int_0^{\bar t}\mu'_\eta(s)\exp((s-t)/\eta)\,ds}{p_\eta(\bar t)+p_\eta^2(\bar t)\mu_\eta(\bar t)}. \tag{6.29}$$

Recalling (6.26) and the assumption $p_\eta(\bar t)>1$, we obtain

$$\mu'_\eta(\bar t)\geqq C+\frac{1}{1+\mu_\eta(\bar t)}\mu'_\eta(\bar t). \tag{6.30}$$

Using Lemma 5.4, it follows that if $p_\eta(\bar t)>1$,

$$\mu'_\eta(\bar t)\geqq C\frac{\exp(-c_1)}{1+\exp(-c_1)}. \tag{6.31}$$

Recalling (6.28), we obtain (6.25), as needed.

(d) We can now prove Lemma 6.4.

LEMMA 6.4. *There exist two constants c_6, $c_7>0$ (independent of $\eta\in\,]0,+\infty[$) such that*

$$p'_\eta(t)\geqq -c_6, \qquad t\in[0,T], \tag{6.32}$$

$$\frac{1-p_\eta(t)}{\eta}\geqq -c_7, \qquad t\in\,]0,T[. \tag{6.33}$$

Proof. Recalling the relation (6.9) and using the estimates already proved, we readily obtain the relation (6.32). The proof of (6.33) is similar, but starting from the relation (6.5).

LEMMA 6.5. *We have that*

$$p_\eta \to 1 \quad \textit{uniformly in } [0, T]. \tag{6.34}$$

Proof. Let

$$q_\eta(t) = \begin{cases} p_\eta(t) - 1 + c_6 t, & t \in [0, T], \\ \dfrac{\exp 2c_1 - p_\eta(T)}{T}(t - T) + p_\eta(T) - 1 + c_6 T, & t \in [T, 2T] \end{cases} \tag{6.35}$$

where c_1 and c_6 are the constants appearing in (5.21) and (6.32). Thanks to Lemmas 5.4 and 6.4, $\{q_\eta\}$ is an equicontinuous family of nondecreasing functions defined in $[0, 2T]$, which verify

$$q_\eta(0) = 0, \qquad q_\eta(2T) = \exp 2c_1 - 1 + c_6 T. \tag{6.36}$$

Looking at the graph of q_η with respect to a new couple of coordinate axes rotated by $\pi/4$ (in the positive sense), we obtain a new family of functions $\{\lambda_\eta\}$ that verify (ρ independent of η)

$$\lambda_\eta : [0, \rho] \to \mathbf{R}, \tag{6.37}$$

$$\{\lambda_\eta\} \text{ are uniformly bounded,} \tag{6.38}$$

$$\lambda_\eta \text{ are equi-Lipschitz (with Lipschitz constant } \leqq 1). \tag{6.39}$$

By the Ascoli Theorem, for every sequence λ_{η_n} ($\eta_n \downarrow 0$) there exists a subsequence $\lambda_{\eta_{n_k}}$, such that

$$\lambda_{\eta_{n_k}} \to \bar{\lambda} \quad \text{uniformly in } [0, \rho], \tag{6.40}$$

where $\bar{\lambda}$ is a bounded and Lipschitz function (with Lipschitz constant $\leqq 1$). Going back to the original axes, the function $\bar{\lambda}$ becomes a bounded (not necessarily continuous) nondecreasing function λ. Moreover we have that

$$q_{\eta_{n_k}}(t) \to \lambda(t) \tag{6.41}$$

for each $t \in [0, 2T]$ where λ is continuous. Since λ is a monotone function, it follows that

$$q_{\eta_{n_k}} \to \lambda \quad \text{a.e. in } [0, 2T]. \tag{6.42}$$

Recalling the definition (6.35) of q_η, we obtain

$$p_{\eta_{n_k}} \to \lambda + 1 - c_6 t \quad \text{a.e. in } [0, T]. \tag{6.43}$$

By Lemma 5.4, we have that $\lambda(t) + 1 - c_6 t > 0$ (almost everywhere in $[0, T]$). Hence,

$$\frac{1}{p_{\eta_{n_k}}(t)} \to \frac{1}{\lambda(t) + 1 - c_6 t} \quad \text{a.e. in } [0, T]. \tag{6.44}$$

Comparing with (6.15), it follows that

$$p_\eta \to 1 \quad \text{a.e. in } [0, T]. \tag{6.45}$$

Since the limit function is continuous we obtain easily that the convergence is uniform in $[0, T]$.

(e) Lemma 6.4 allows us to prove the following estimate:

LEMMA 6.6. *There exist two constants c_8, $c_9 > 0$ (independent of η) such that*

$$z'_\eta(t) \geqq -c_8, \quad t \in [0, T], \quad \eta \in [0, c_9]. \tag{6.46}$$

Proof. If $\eta = 0$ the proof is obvious. As in the proof of Lemma 6.3, we can equivalently prove that

$$\mu'_\eta(t) \leqq c'_8, \qquad t \in [0, T], \tag{6.47}$$

where $\mu_\eta = \exp(-z_\eta)$. Given $\eta \in]0, +\infty[$, let $\bar{t} \in [0, T]$, verifying

$$\mu'_\eta(\bar{t}) \geqq \mu'_\eta(t), \qquad t \in [0, T]. \tag{6.48}$$

The value $\bar{t}$ may depend on η. Obviously we can assume that $\mu'_\eta(\bar{t}) \geqq 0$. Recalling (6.14), we have

$$\mu'_\eta(\bar{t}) \leqq C + \frac{1}{p_\eta(\bar{t}) + p_\eta^2(\bar{t})\mu_\eta(\bar{t})}\mu'_\eta(\bar{t}). \tag{6.49}$$

Hence, by Lemma 5.4,

$$\mu'_\eta(\bar{t}) \leqq C + \frac{1}{p_\eta(\bar{t}) + p_\eta^2(\bar{t})\exp(-c_1)}\mu'(\bar{t}). \tag{6.50}$$

That is

$$\mu'_\eta(\bar{t})B_\eta \leqq C, \tag{6.51}$$

where

$$B_\eta = 1 - \frac{1}{p_\eta(\bar{t}) + p_\eta^2(\bar{t})\exp(-c_1)}. \tag{6.52}$$

By Lemma 6.5

$$\lim_{\eta \downarrow 0} B_\eta = \frac{\exp(-c_1)}{1 + \exp(-c_1)}. \tag{6.53}$$

Hence there exists $c_9 > 0$, such that

$$B_\eta \geqq \frac{1}{2}\frac{\exp(-c_1)}{1+\exp(-c_1)}, \qquad \eta \in]0, c_9[, \tag{6.54}$$

that is,

$$\mu'_\eta(\bar{t}) \leqq 2C\frac{1+\exp(-c_1)}{\exp(-c_1)}. \tag{6.55}$$

(f) The following result can now be obtained.

LEMMA 6.7. *There exists a constant* $c > 0$ (*independent of* $\eta \in]0, +\infty[$) *such that* (*for all* $t \in [0, T]$)

$$|z'_\eta(t)| \leqq c, \tag{6.56}$$

$$|p'_\eta(t)| \leqq c, \tag{6.57}$$

$$\frac{1}{\eta}|1 - p_\eta(t)| \leqq c. \tag{6.58}$$

Proof. The relation (6.56) is a consequence of Lemmas 6.2, 6.3, and 6.6. The relations (6.57) and (6.58) are implied by (6.56) and the relations (6.9) and (6.5).

(g) At last we are now able to prove a result on the asymptotic behavior of the solution for $\eta \to +\infty$ and for $\eta \downarrow 0$.

THEOREM 6.1. *We have that*

$$\lim_{\eta \downarrow 0} z_\eta = z_0 \quad \textit{uniformly on } [0, T], \tag{6.59}$$

$$\lim_{\eta \to +\infty} z_\eta = z_\infty \quad \textit{uniformly on } [0, T], \tag{6.60}$$

where z_0 (*respectively*, z_η, z_∞), *is the solution of Problem* 4.2 *for* $\eta = 0$ (*respectively*, $\eta \in]0, +\infty[$, $\eta = +\infty$).

Proof. We begin proving (6.59). Let η_n be a sequence such that $\eta_n \downarrow 0$. By Lemmas 6.5 and 6.7, there exist a subsequence η_{n_k} and $\underline{z} \in C^0([0, T])$ such that

$$\lim_{k \to +\infty} p_{\eta_{n_k}} = 1 \quad \text{uniformly in } [0, T], \tag{6.61}$$

$$\lim_{k \to +\infty} z_{\eta_{n_k}} = \underline{z} \quad \text{uniformly in } [0, T], \tag{6.62}$$

$$\lim_{k \to +\infty} p'_{\eta_{n_k}} = 0 \quad \text{weakly in } L^2(]0, T[), \tag{6.63}$$

$$\lim_{k \to +\infty} z'_{\eta_{n_k}} = \underline{z}' \quad \text{weakly in } L^2(]0, T[). \tag{6.64}$$

Now let ($\eta \in [0, +\infty]$)

$$\alpha_\eta(t) = \frac{1}{p_\eta^2(t)[\exp(-z_\eta(t)) + \int_{\mathbf{R}} (Uz_\eta)(x, t)\, dx}. \tag{6.65}$$

By (6.61) and (6.62) we can prove easily that

$$\lim_{k \to +\infty} \alpha_{\eta_k} = \frac{1}{\exp(-\underline{z}) + \int_{\mathbf{R}} U\underline{z}\, dx} \quad \text{uniformly in } [0, T]; \tag{6.66}$$

hence,

$$\lim_{k \to +\infty} p'_{\eta_{n_k}} \alpha_{\eta_{n_k}} = 0 \tag{6.67}$$

in $\mathcal{D}'(]0, T[)$ (for instance).

Passing to the limit (as $k \to +\infty$) in the relation (6.9), it follows that

$$\underline{z}' = -\frac{A\underline{z}}{\exp(-\underline{z}) + \int_{\mathbf{R}} (U\underline{z})(x, t)\, dx}, \tag{6.68}$$

which implies that $\underline{z} \in C^1([0, T])$. This means that $\underline{z} = z_0$, where z_0 is the unique solution of Problem 4.2 with $\eta = 0$. Since the limit function $\underline{z}$ is independent of the sequence η_n, the relation (6.59) follows. We conclude by proving (6.60). Let η_n be a sequence such that $\eta_n \to +\infty$. By Lemmas 6.1 and 6.7, there exists a subsequence η_{n_k} and $\bar{z} \in C^0([0, T])$ such that

$$\lim_{k \to +\infty} z_{\eta_{n_k}} = \bar{z} \quad \text{uniformly in } [0, T], \tag{6.69}$$

$$\lim_{k \to +\infty} p_{\eta_{n_k}} = \exp(\bar{z}) \quad \text{uniformly in } [0, T], \tag{6.70}$$

$$\lim_{k \to +\infty} z'_{\eta_{n_k}} = \bar{z}' \quad \text{weakly in } L^2(]0, T[). \tag{6.71}$$

Recalling Lemma 5.4 and passing to the limit in the relation (6.5), we have

$$\bar{z}' = -\frac{A\bar{z}}{2\exp(-\bar{z}) + \int_{\mathbf{R}} (U\bar{z})(x, t)\, dx}. \tag{6.72}$$

As in the case $\eta \downarrow 0$, we can now obtain (6.60).

7. Numerical approximation. The equivalent formulation of Problem 2.1 in Problem 4.2 suggests a first way to approximate the solution. It comes down to numerically solve an integrodifferential equation along a characteristic line. This is a rather classical task.

Another approach, which we have used to obtain numerical simulations in [8], consists in solving Problem 2.1 directly in an implicit way at each time level. The underlying idea is like the one introduced in [17] and next adapted in [5], [7], [9], [13] for solving models with elastic passive elements. After a discretization of the time t with step Δt, we suppose to know the approximated values $\bar{n}(x, t)$ for $x \in \mathbf{R}$ and we want to compute $\bar{n}(x, t+\Delta t)$. This is achieved in two successive steps.

Step 1. Neglecting the interfilament notion, we solve the equation

$$\frac{dn(x, t)}{dt} = f(x)\gamma(t)(1 - n(x, t)) - g(x)n(x, t) \quad \text{from } t \text{ to } t+\Delta t. \tag{7.1}$$

We denote $\bar{n}_*(x, t+\Delta t)$ the solution, which is, therefore, the result in $(t, t+\Delta t)$ only of the chemical processes of association and dissociation of bridges.

We denote by $\overline{FCE}_*(t+\Delta t)$ the corresponding contractile element force, that is,

$$\overline{FCE}_*(t+\Delta t) = k \int_{-\infty}^{+\infty} \bar{n}_*(\xi, t+\Delta t)\xi\, d\xi. \tag{7.2}$$

Now, in order to restore the equilibrium of the contractile and series forces we must suitably translate $\bar{n}_*(x, t+dt)$ (interfilament motion).

Step 2. We set

$$\bar{n}(x, t+\Delta t) = \bar{n}_*(x+\delta, t+\Delta t), \qquad x \in \mathbf{R}.$$

The shift number δ is computed from (2.9) in the following way. We denote, for the sake of brevity, by $\bar{\sigma}_\delta(t+\Delta t)$ the contractile force related to the translated function $\bar{n}_*(x+\delta, t+\Delta t)$, that is,

$$\bar{\sigma}_\delta(t+\Delta t) = \int_{-\infty}^{+\infty} \bar{n}_*(\xi+\delta, t+\Delta t)\xi\, d\xi. \tag{7.3}$$

Note that as $\bar{n}_*$ vanishes for $|x| \to +\infty$ we have

$$\bar{\sigma}_\delta(t+\Delta t) = \sigma_0(t+\Delta t) - \delta KCE_* \tag{7.4}$$

where

$$KCE_* = \int_{-\infty}^{+\infty} \bar{n}_*(\xi, t+\Delta t)\, d\xi, \qquad \sigma_0 = \overline{FCE}_*(t+\Delta t).$$

We discretize now (2.9) by means of an implicit scheme by setting

$$\frac{d\sigma(t+\Delta t)}{dt} \simeq \frac{(\bar{\sigma}_\delta(t+\Delta t) - \bar{\sigma}(t))}{\Delta t},$$

$$\frac{d\varepsilon(t+\Delta t)}{dt} \simeq \frac{(\varepsilon(t+\Delta t) - \varepsilon(t))}{\Delta t}.$$

Note that $\varepsilon(t+\Delta t)=\varepsilon(t)+\delta$. The shift parameter δ is then the solution of the following nonlinear equation:

$$\frac{\bar{\sigma}_\delta(t+\Delta t)-\bar{\sigma}(t)}{\Delta t}=\frac{\delta}{\Delta t}(2ab+b\bar{\sigma}_\delta(t+\Delta t))$$

$$-\frac{b}{\eta}(\bar{\sigma}_\delta(t+\Delta t)-a(\exp(b(\varepsilon(t)+\delta-\varepsilon_0))-1)) \tag{7.5}$$

$$\cdot(a+\bar{\sigma}_\delta(t+\Delta t)-a(\exp(b(\varepsilon(t)+\delta-\varepsilon_0))-1)).$$

There exist two solutions of (7.5) one of which is spurious because of the term $(a+\sigma-\sigma_p)$; this one could be eliminated by computing the last term in t instead of $t+\Delta t$ (semi-implicit method). The good solution is that which in the limit case $\eta=0$ gives: $\sigma(t+\Delta t)=\sigma_p(t+\Delta t)$, the equilibrium between contractile and series forces. This value can be in practice easily computed by means of Newton method.

Adapting the analysis developed in [9] and [13] for the elastic passive elements models, it is possible to prove that for $\Delta t\to 0$ we have the convergence of $\bar{n}(x, t)$ to $n(x, t)$ in particular the uniform convergence of $\overline{FCE}(t)$ to FCE in $[0, T]$, $T>0$ fixed.

REFERENCES

[1] F. BROWDER, *Problémes non linéaires*, Lecture Notes University of Montreal, Canada, 1966.

[2] F. BUCHTAL AND E. KAISER, *The rheology of the cross-striated muscle fibre with special reference to isotonic conditions*, Dan. Biol. Med., 21 (1951), pp. 7-123.

[3] A. CAPELO, V. COMINCIOLI, R. MINELLI, C. POGGESI, C. REGGIANI, AND L. RICCIARDI, *Study and parameters identification of a rheological model for excised quiescent cardiac muscle*, J. Biomechanics, 14 (1981), pp. 1-11.

[4] P. L. COLLI, *A mathematical model of heterogeneous behavior of single muscle fibres*, J. Math. Biol., 24 (1986), pp. 661-683.

[5] V. COMINCIOLI AND A. TORELLI, *Mathematical aspects of the cross-bridge mechanism in muscle contraction*, Nonlinear Anal., Theory, Methods Appl., 7 (1983), pp. 661-683.

[6] V. COMINCIOLI, C. POGGESI, C. REGGIANI, AND A. TORELLI, *Mathematical models for contracting muscle*, in Numerical Solutions of Nonlinear Problems, France-Italy-USSR 6 Joint Symp., INRIA Rocquencourt, France, 1983 (1984).

[7] V. COMINCIOLI, C. POGGESI, AND A. TORELLI, *A four-state cross-bridge model for muscle contraction. Mathematical study and validation*, J. Math. Biol., 20 (1984), pp. 277-304.

[8] V. COMINCIOLI, R. BOTTINELLI, R. MINELLI, C. POGGESI, C. REGGIANI, L. RICCIARDI, AND A. TORELLI, *Mathematical model of contracting muscle*, HUSPI 10 (1985).

[9] J. DOUGLAS AND F. A. MILNER, *Numerical methods for a model of cardiac muscle contraction*, Calcolo, XX (1983), pp. 123-141.

[10] L. GASTALDI AND F. TOMARELLI, *A nonlinear hyperbolic Cauchy problem arising in the dynamic of cardiac muscle*, Pubb. I.A.N. of C.N.R. Pavia, 340 (1983), pp. 1-24.

[11] ———, *A uniqueness result for a nonlinear hyperbolic equation*, Annali di Matematica Pura ed Applicata (IV), 137 (1984), pp. 175-205.

[12] ———, *A nonlinear and nonlocal equation describing the muscle contraction*, Nonlinear Anal., Theory, Methods Appl., 10 (1987), pp. 163-182.

[13] ———, *Numerical approximation of nonlinear problem arising in physiology*, Pubb. I.A.N. of C.N.R. Pavia, 429 (1984), pp. 1-17.

[14] A. S. GLANTZ, *A three-element description for muscle with viscoelastic passive elements*, J. Biomechanics, 10 (1977), pp. 5-20.

[15] A. V. HILL, *Abrupt transition from rest to activity in muscle*, Proc. Roy. Soc. London Ser. B, 136 (1949), pp. 399-420.

[16] A. F. HUXLEY, *Muscle structure and theories of contraction*, Progr. Biophys., 7 (1957), pp. 255-318.

[17] F. J. JULIAN, *Activation in a skeletal muscle contraction model with a modification for insect fibriller muscle*, Biophys. J., 9 (1969), pp. 547-570.

[18] R. MINELLI, C. REGGIANI, R. DIONIGI, AND V. CAPPELLI, *Cardiac muscle models for both isotonic and isometric contractions*, Pflugers Arch., 359 (1975), pp. 69-80.

[19] M. I. M. NOBLE, *The diastolic viscous properties of cat papillary muscle*, Circulation Res., 40 (1977), pp. 288-292.

[20] R. B. PANERAI, *A model of cardiac muscle mechanics and energetics*, J. Biomechanics, 13 (1980), pp. 929-940.

[21] G. H. POLLACK, *The cross-bridge theory*, Physiological Rev., 63 (1983), pp. 1049-1113.

[22] C. REGGIANI, C. POGGESI, AND R. MINELLI, *The analysis of some mechanical properties of quiescent myocardium*, J. Biomechanics, 12 (1979), pp. 173-182.

[23] C. REGGIANI, V. COMINCIOLI, V. CAPPELLI, AND C. POGGESI, *Age dependent changes in time course and load sensitivity of the relaxation phase in rat cardiac muscle*, to appear.

[24] A. Y. K. WONG, *Mechanics of cardiac muscle, based on Huxley's model: mathematical simulation of isometric contraction*, J. Biomechanics, 4 (1971), pp. 529-540.

[25] ———, *A model of excitation-contraction coupling in frog cardiac muscle*, J. Biomechanics, 9 (1976), pp. 319-332.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

008

BOUNDARY INTEGRAL OPERATORS ON LIPSCHITZ DOMAINS: ELEMENTARY RESULTS*

MARTIN COSTABEL†

Abstract. The simple and double layer potentials for second order linear strongly elliptic differential operators on Lipschitz domains are studied and it is shown that in a certain range of Sobolev spaces, results on continuity and regularity can be obtained without using either Calderón's theorem on the L_2-continuity of the Cauchy integral on Lipschitz curves [J. L. Journé, "Calderón-Zygmuno operators, pseudo-differential operators and the Cauchy integral of Calderón," in *Lecture Notes in Math.* 994, Springer-Verlag, Berlin, 1983] or Dahlberg's estimates of harmonic measures ["On the Poisson integral for Lipschitz and C^1 domains," *Studia Math.*, 66 (1979), pp. 7-24]. The operator of the simple layer potential and of the normal derivative of the double layer potential are shown to be strongly elliptic in the sense that they satisfy Gårding inequalities in the respective energy norms. As an application, error estimates for Galerkin approximation schemes for integral equations of the first kind are derived.

Key words. Lipschitz domains, layer potentials, trace lemma, jump relations, Green's formula, Galerkin approximation, Gårding's inequality

AMS(MOS) subject classifications. 45E99, 35J25, 45L10, 58G15

1. Introduction.

Boundary value problems on Lipschitz domains and the method of layer potentials for their solution have attracted some attention in recent years both in the theoretical and the applied mathematical literature.

On one hand, the final proof by Coifman, McIntosh, and Meyer [5] of Calderón's Theorem on the L_2-continuity of the Cauchy integral on Lipschitz curves and Dahlberg's estimates [10] of the Poisson kernel paved the way for investigations of the Dirichlet and Neumann problems for the Laplace equation by means of boundary integral equations [12], [21], [27]. This method was also applied to some boundary value problems for the equations of linear elasticity theory [20].

On the other hand, in the applied sciences, the so-called boundary element methods are frequently used for domains with corners and edges without mathematical analysis being available. As long as there exists no elementary proof of Calderón's theorem and its consequences, it seems justified to study the range of possible results obtainable without this deep and, for the nonspecialist, not easily accessible result.

We use throughout the weak (distributional) definition of boundary values and show that the operators of the simple layer, the double layer, the normal derivative of the simple layer, and the normal derivative of the double layer define bounded operators in those Sobolev spaces on the boundary that correspond to the "energy norm," i.e., to the variational formulation of the boundary value problem. The simple layer and the normal derivative of the double layer define strongly elliptic operators. This implies stability of corresponding Galerkin approximation schemes.

In order to show continuity of the operators in a certain range of Sobolev spaces, we prove a generalization of Gagliardo's Trace Lemma (Lemma 3.6) and use regularity results for the Dirichlet and Neumann problems by Nečas [23]. Nečas obtained these results by elementary means, applying an identity of Rellich that had been used for similar purposes by Payne and Weinberger [24] and recently by Jerison and Kenig [16], [17] and Verchota [21], [27]. The same tools yield regularity results for the solutions of the integral equations and also invertibility under some hypotheses on the

* Received by the editors December 16, 1985; accepted for publication (in revised form) June 18, 1987.

† Mathematisches Institut, Technische Hochschule Darmstadt, D-6100 Darmstadt, Federal Republic of Germany.

differential operator and its fundamental solution satisfied for instance in the case of the Laplace operator unless the domain is a subset of $\mathbb{R}^2$ with analytic capacity equal to one.

This work is part of the author's habilitation thesis [7]. Further results concerning strong ellipticity of boundary integral operators for higher order differential equations on smooth domains have been published [9]. That paper also contains an extensive list of references on the Calderón-Seeley-Hörmander method of boundary integral equations for elliptic boundary value problems and on the history of strong ellipticity for boundary integral operators. Let us mention here only two references from each of these two fields: The books by Chazarain and Piriou [4] and by Dieudonné [11] describe the method of the Calderón projector of elliptic equations of any order on smooth domains. The lecture notes by Nedelec [22] and the paper by Hsiao and Wendland [15] contain, for the example of the Laplace operator on smooth domains, the idea of transforming the strong ellipticity of the differential operator via Green's formula into the strong ellipticity of certain operators on the boundary (see the proof of Theorem 2 below).

2. Main results. In this section we state the main results of this paper. Proofs are given in § 4.

Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain. This means that its boundary Γ is locally the graph of a Lipschitz function. For properties of Lipschitz domains we refer to Nečas [23] and Grisvard [14]. Because of the invariance of the Sobolev spaces $H^s = W^{s,2}$ under Lipschitz coordinate transformations for $|s| \leqq 1$, we can define the spaces $H^s(\Gamma)(|s| \leqq 1)$ in the usual way using local coordinate representations of the Lipschitz manifold Γ. The same reason implies Gagliardo's Trace Lemma:

$$\gamma_0 \colon u \mapsto \gamma_0 u := u|_\Gamma \colon H^s_{\mathrm{loc}}(\mathbb{R}^n) \to H^{s-1/2}(\Gamma) \text{ is continuous}$$

(2.1) and has a continuous right inverse

$$\gamma_0^- \colon H^{s-1/2}(\Gamma) \to H^s_{\mathrm{loc}}(\mathbb{R}^n) \quad \text{for all } s \in (\tfrac{1}{2}, 1].$$

Here traces are understood in the distributional sense, i.e., the mapping γ_0 is well defined for smooth (say continuous) functions, and for arbitrary $u \in H^s_{\mathrm{loc}}(\mathbb{R}^n)$ it is defined by approximating u by smooth functions.

Let

$$P = -\sum_{j,k=1}^{n} \partial_j a_{jk}\, \partial_k + \sum_{j=1}^{n} b_j\, \partial_j + c \tag{2.2}$$

be a differential operator with $C^\infty(\mathbb{R}^n; \mathbb{C})$-coefficients a_{jk}, b_j and c. Here $\partial_j = \partial/\partial x_j$.

We emphasize that all results will also be valid in the case of systems, i.e., for matrix valued coefficients a_{jk}, b_j and c. It is only for notational convenience that we stick to the scalar case.

We assume that P is strongly elliptic which implies that for the bilinear form

$$\Phi_\Omega(u, v) := \int_\Omega \left(\sum_{j,k=1}^{n} a_{jk}\, \partial_k u\, \overline{\partial_j v} + \sum_{j=1}^{n} b_j\, \partial_j u \bar{v} + c u \bar{v} \right) dx \tag{2.3}$$

there holds a Gårding inequality on all of $H^1(\Omega)$

$$\operatorname{Re} \Phi_\Omega(u, u) \geqq \lambda \|u\|^2_{H^1(\Omega)} - C \|u\|^2_{L_2(\Omega)} \quad \text{for all } u \in H^1(\Omega) \tag{2.4}$$

with some $\lambda > 0$. (In the case of systems we have to require (2.4) explicitly. It holds, for example, for the equations of linear elasticity theory by virtue of Korn's inequality [23, p. 194].)

Furthermore we assume that P has a fundamental solution G that is a two-sided inverse of P on the space of compactly supported distributions on $\mathbb{R}^n$. Then G has a weakly singular kernel that we also denote by G, and the function $(x, y) \mapsto G(x, y)$ is C^∞ outside the diagonal of $\mathbb{R}^n \times \mathbb{R}^n$.

For a locally integrable function v on Γ we then can define the simple layer potential

$$K_0 v(x) := \int_\Gamma G(x, y) v(y)\, ds(y) \qquad (x \in \mathbb{R}^n \backslash \Gamma) \tag{2.5}$$

where ds is the surface measure on Γ, and the double layer potential

$$K_1 v(x) := \int_\Gamma \partial_{\nu(y)} G(x, y) v(y)\, ds(y). \tag{2.6}$$

Here ∂_ν is the conormal derivative

$$\partial_\nu := \sum_{j,k=1}^{n} n_j a_{jk}\, \partial_k, \tag{2.7}$$

where n_j are the components of the almost everywhere defined outward pointing normal vector.

The boundary integral operators are defined by taking the boundary data of K_0 and K_1 (in the distributional sense; see (2.1) and Lemma 3.2 below)

$$\begin{aligned} Av &:= \gamma_0 K_0 v, & Bv &:= \gamma_1 (K_0 v|_\Omega), \\ Cv &:= \gamma_0 (K_1 v|_\Omega), & Dv &:= -\tilde{\gamma}_1 (K_1 v|_\Omega). \end{aligned} \tag{2.8}$$

Here $\gamma_1 u := \partial_\nu u|_\Gamma$ and $\tilde{\gamma}_1 u := \partial_\nu u|_\Gamma - \sum_{j=1}^n n_j b_j u|_\Gamma$.

Under these assumptions, we have the following continuity result.

THEOREM 1. *For all $\sigma \in (-\frac{1}{2}, \frac{1}{2})$ the following operators are continuous*:

(i) $K_0 : H^{-1/2+\sigma}(\Gamma) \to H^{1+\sigma}_{\mathrm{loc}}(\mathbb{R}^n)$;
(ii) $K_1 : H^{1/2+\sigma}(\Gamma) \to H^{1+\sigma}(\Omega)$;
(iii) $A : H^{-1/2+\sigma}(\Gamma) \to H^{1/2+\sigma}(\Gamma)$;
(iv) $B : H^{-1/2+\sigma}(\Gamma) \to H^{-1/2+\sigma}(\Gamma)$;
(v) $C : H^{1/2+\sigma}(\Gamma) \to H^{1/2+\sigma}(\Gamma)$;
(vi) $D : H^{1/2+\sigma}(\Gamma) \to H^{-1/2+\sigma}(\Gamma)$.

Remark. As shown by Verchota [27] and Jerison and Kenig [16], [17], the Calderón and Dahlberg theorems give the above results for the endpoint $\sigma = \frac{1}{2}$. An argument using duality and interpolation then allows to cover the whole range $\sigma \in [-\frac{1}{2}, \frac{1}{2}]$, which is optimal in the sense that, for Lipschitz boundaries, Sobolev spaces $H^s(\Gamma)$ with $|\sigma| > 1$ cannot be defined in a unique invariant way.

The operators A and D are strongly elliptic.

THEOREM 2. *There exist compact operators*

$$T_A : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma), \qquad T_D : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$$

and constants $\lambda_A, \lambda_D > 0$ such that

$$\mathrm{Re}\langle (A + T_A) v, \bar{v} \rangle \geqq \lambda_A \|v\|^2_{H^{-1/2}(\Gamma)} \quad \textit{for all } v \in H^{-1/2}(\Gamma), \tag{2.9}$$

$$\mathrm{Re}\langle (D + T_D) v, \bar{v} \rangle \geqq \lambda_D \|v\|^2_{H^{1/2}(\Gamma)} \quad \textit{for all } v \in H^{1/2}(\Gamma). \tag{2.10}$$

Here the brackets $\langle \cdot, \cdot \rangle$ denote the natural duality between a Sobolev space $H^s(\Gamma)$ and its dual $H^{-s}(\Gamma)$.

The following regularity result holds.

THEOREM 3. *Let $\sigma \in [0, \frac{1}{2}]$ and let $\psi \in H^{-1/2}(\Gamma)$ and $v \in H^{1/2}(\Gamma)$ satisfy*

$$A\psi \in H^{1/2+\sigma}(\Gamma) \quad \textit{or} \quad B\psi \in H^{-1/2+\sigma}(\Gamma)$$

and

$$Cv \in H^{1/2+\sigma}(\Gamma) \quad \text{or} \quad Dv \in H^{-1/2+\sigma}(\Gamma).$$

Then $\psi \in H^{-1/2+\sigma}(\Gamma)$ *and* $v \in H^{1/2+\sigma}(\Gamma)$, *and there hold the a priori estimates*

$$\|\psi\|_{H^{-1/2+\sigma}(\Gamma)} \leqq C(\|A\psi\|_{H^{1/2+\sigma}(\Gamma)} + \|\psi\|_{H^{-1/2}(\Gamma)}), \tag{2.11}$$

$$\|\psi\|_{H^{-1/2+\sigma}(\Gamma)} \leqq C(\|B\psi\|_{H^{-1/2+\sigma}(\Gamma)} + \|\psi\|_{H^{-1/2}(\Gamma)}), \tag{2.12}$$

$$\|v\|_{H^{1/2+\sigma}(\Gamma)} \leqq C(\|Cv\|_{H^{1/2+\sigma}(\Gamma)} + \|v\|_{H^{1/2}(\Gamma)}), \tag{2.13}$$

$$\|v\|_{H^{1/2+\sigma}(\Gamma)} \leqq C(\|Dv\|_{H^{-1/2+\sigma}(\Gamma)} + \|v\|_{H^{1/2}(\Gamma)}). \tag{2.14}$$

Now let $(S^h)_{h>0}$ be a family of subspaces of $H^{-1/2}(\Gamma)$ with the property that the orthogonal projection operators onto S^h tend strongly to the identity in $H^{-1/2}(\Gamma)$ for $h \to 0$.

For the equation

$$Av = g \quad \text{with } g \in H^{1/2}(\Gamma) \tag{2.15}$$

we consider the Galerkin scheme

$$\text{Find } v_h \in S^h \text{ such that } \langle Av_h, w\rangle = \langle g, w\rangle \text{ for all } w \in S^h. \tag{2.16}$$

From Theorem 2 then follows stability and convergence in $H^{-1/2}(\Gamma)$. Note that Theorem 2 implies that the operators A and D are Fredholm operators of index zero.

THEOREM 4. *If the operator* $A: H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ *is injective then for any* $g \in H^{1/2}(\Gamma)$ *there is a* $h_0 > 0$ *such that for all* $0 < h < h_0$ *the Galerkin scheme* (2.16) *has a unique solution* $v_h \in S^h$. *For* $h \to 0$, v_h *converges to the unique solution* $v \in H^{-1/2}(\Gamma)$ *of* (2.15) *quasioptimally, i.e., there exists a constant* C *such that for all* $0 < h < h_0$

$$\|v - v_h\|_{H^{-1/2}(\Gamma)} \leqq C \inf_{w \in S^h} \|v - w\|_{H^{-1/2}(\Gamma)}. \tag{2.17}$$

Of course, a corresponding result holds for the operator D.

From Theorem 3 we can deduce asymptotic error estimates using (2.17). We assume for instance that S^h are regular finite element spaces, in the simplest case, e.g., consisting of functions piecewise constant on Γ that are constant on the faces of a triangulation of Γ quasiuniform with respect to h where h is the meshsize. Then there holds the following theorem.

THEOREM 5. *Let* A *be injective as above and* $g \in H^1(\Gamma)$. *Then there is a constant* C *such that for all* $0 < h < h_0$

$$\|v - v_h\|_{H^{-1/2}(\Gamma)} \leqq Ch^{1/2}\|g\|_{H^1(\Gamma)}. \tag{2.18}$$

3. The tools. In this section we collect some results, some new but most of them known, and adapt them to the present situation.

We need some further notation.

$H^s_{\text{comp}}(\mathbb{R}^n)$ is the space of distributions in H^s with compact support. It is thus in a natural way the dual space of $H^{-s}_{\text{loc}}(\mathbb{R}^n)$.

$$H^s_P(\Omega) := \{u \in H^s(\Omega) \mid Pu \in L_2(\Omega)\}, \qquad \|u\|^2_{H^s_P(\Omega)} := \|u\|^2_{H^s} + \|Pu\|^2_{L_2},$$

$$P' := -\sum_{j,k=1}^{n} \partial_j a_{jk} \partial_k - \sum_{j=1}^{n} \partial_j b_j + c \quad \text{is the formal transpose of } P.$$

We may assume $a_{jk} = a_{kj}$ without restriction.

By P^0 we denote any operator with the same principal part as P and positive on $H^1(\Omega)$. We may take, e.g.,

$$-\sum_{j,k=1}^{n} \partial_j a_{jk} \partial_k + \lambda \quad \text{with } \lambda > 0.$$

Thus there holds with some $\lambda > 0$

$$\text{Re}(P^0 u, u)_{L_2(\Omega)} \geqq \lambda \|u\|^2_{H^1(\Omega)} \quad \text{for all } u \in C^2(\bar{\Omega}). \tag{3.1}$$

It follows with the trace lemma (2.1) that for P^0 the Dirichlet problem is uniquely solvable in the weak sense.

LEMMA 3.1. *The Dirichlet problem*

$$P^0 u = 0 \quad \text{in } \Omega, \qquad \gamma_0 u = v$$

has for $v \in H^{1/2}(\Gamma)$ *a unique solution* $u := Tv \in H^1(\Omega)$. *The solution operator* $T: H^{1/2}(\Gamma) \to H^1_P(\Omega)$ *is continuous.*

From the partial integration formula

$$\int_\Omega (\partial_j u v + u\, \partial_j v)\, dx = \int_\Gamma u v n_j\, ds \quad \text{for } u, v \in H^1(\Omega)$$

follow [23] the *first Green formula*

$$\int_\Omega \bar{v} P u\, dx = \Phi_\Omega(u, v) - \int_\Gamma \partial_\nu u \bar{v}\, ds \quad \text{for } v \in H^1(\Omega), \quad u \in H^2(\Omega) \tag{3.2}$$

and the *second Green formula*

$$\int_\Omega (uP'v - vPu)\, dx = \int_\Gamma (v\, \partial_{\tilde{\nu}} u - u\, \partial_{\tilde{\nu}} v)\, ds \quad \text{for } u, v \in H^2(\Omega) \tag{3.3}$$

where we defined

$$\partial_{\tilde{\nu}} u := \partial_\nu u - \sum_{j=1}^{n} n_j b_j u.$$

Now let u be a function defined on $\mathbb{R}^n$ such that

$$u_1 := u|_\Omega \in C^\infty(\bar{\Omega}) \quad \text{and} \quad u_2 := u|_{\Omega^c} \in C^\infty_{\text{comp}}(\bar{\Omega}^c),$$

where $\Omega^c := \mathbb{R}^n \backslash \bar{\Omega}$ is the exterior domain. Let $f := Pu|_{\mathbb{R}^n \backslash \Gamma}$ and let

$$[u] := \gamma_0 u_2 - \gamma_0 u_1 \quad \text{denote the jump of } u \text{ across } \Gamma.$$

Then there holds the *representation formula* (for $x \in \mathbb{R}^n \backslash \Gamma$)

$$u(x) = Gf(x) + \int_\Gamma (\partial_{\nu(y)} G(x, y)[u(y)] - G(x, y)[\partial_{\tilde{\nu}(y)} u(y)])\, ds(y). \tag{3.4}$$

We shall need equations (3.2)–(3.4) for more general functions. To this purpose, we first define the conormal derivative in the weak sense by using the first Green formula (3.2). Recall γ_0^- from (2.1).

LEMMA 3.2. *Let* $u \in H^1_P(\Omega)$. *Then the mapping*

$$\varphi \mapsto \langle \gamma_1 u, \varphi \rangle := \Phi_\Omega(u, \gamma_0^- \varphi) - \int_\Omega Pu \cdot \gamma_0^- \varphi\, dx$$

is a continuous linear functional $\gamma_1 u$ *on* $H^{1/2}(\Gamma)$ *that coincides for* $u \in H^2(\Omega)$ *with the functional defined by* $\partial_\nu u|_\Gamma \in L_2(\Gamma) \subset H^{-1/2}(\Gamma)$.

(3.5) *The mapping* $\gamma_1 : H^1_P(\Omega) \mapsto H^{-1/2}(\Gamma)$ *is continuous.*

The following lemma was shown by Grisvard [14] for the case $P = -\Delta$. The proof works verbatim for the present case.

LEMMA 3.3. $C^\infty(\bar{\Omega})$ *is dense in the Hilbert space* $H^1_P(\Omega)$.

Thus we can extend (3.2)-(3.4) by continuity.

LEMMA 3.4. (i) *The first Green formula in the form*

$$\int_\Omega \bar{v} P u \, dx = \Phi_\Omega(u, v) - \langle \gamma_1 u, \overline{\gamma_0 v} \rangle \tag{3.6}$$

holds for all $u \in H^1_P(\Omega)$, $v \in H^1(\Omega)$.

(ii) *The second Green formula in the form*

$$\int_\Omega (uP'v - vPu) \, dx = \langle \widetilde{\gamma_1} u, \gamma_0 v \rangle - \langle \gamma_1 v, \gamma_0 u \rangle \tag{3.7}$$

holds for all $u, v \in H^1_P(\Omega)$. *Here we define, corresponding to the definition of* $\partial_{\tilde{P}}$:

$$\widetilde{\gamma_1} u := \gamma_1 u - \sum_{j=1}^{n} n_j b_j \gamma_0 u.$$

(iii) *The representation formula in the form*

$$u(x) = Gf(x) + \langle \gamma_1 G(x, \cdot), [\gamma_0 u] \rangle - \langle [\widetilde{\gamma_1} u], G(x, \cdot) \rangle \qquad (x \in \mathbb{R}^n \setminus \Gamma) \tag{3.8}$$

holds for all $u \in L_2(\mathbb{R}^n)$ *with* $u|_\Omega \in H^1(\Omega)$, $u|_{\Omega^c} \in H^1_{\text{comp}}(\Omega^c)$, *and* $f = Pu|_{\mathbb{R}^n \setminus \Gamma} \in L_2(\mathbb{R}^n)$.

The proof is immediate if we keep in mind that $H^1_{P'} = H^1_P(\Omega)$ and that γ_1 remains the same, whether defined from P or from P'. For (3.8) we need only (3.7) and the representation formula (3.4) for a smooth domain, let us say a small ball enclosing the point x.

The following result will be needed in the proof of the jump relations (Lemma 4.1).

LEMMA 3.5. *The trace map*

$$(\gamma_0, \gamma_1) : \varphi \mapsto (\gamma_0 \varphi, \gamma_1 \varphi)$$

maps $C^\infty_{\text{comp}}(\mathbb{R}^n)$ *onto a dense subspace of* $H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$.

Proof. Assume that for some $(\chi, \psi) \in H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma)$ there holds

$$\langle \chi, \gamma_1 \varphi \rangle = \langle \psi, \gamma_0 \varphi \rangle \quad \text{for all } \varphi \in C^\infty_{\text{comp}}(\mathbb{R}^n). \tag{3.9}$$

We have to show that $\chi = \psi = 0$.

Let $T\chi \in H^1_P(\Omega)$ be the solution of the Dirichlet problem (see Lemma 3.1)

$$P^0 T\chi = 0 \quad \text{in } \Omega, \qquad \gamma_0 T\chi = \chi.$$

For arbitrary $f \in L_2(\Omega)$ let $Sf \in H^1_P(\Omega)$ be the unique weak solution of the Dirichlet problem

$$P^0 Sf = f \quad \text{in } \Omega, \qquad \gamma_0 Sf = 0.$$

The second Green formula (3.7) for the operator $P^0 = P^{0'}$ gives

$$\begin{aligned} \langle \gamma_1 Sf, \chi \rangle &= \langle \gamma_1 Sf, \gamma_0 T\chi \rangle - \langle \gamma_1 T\chi, \gamma_0 Sf \rangle \\ &= \int_\Omega (Sf \cdot P^0 T\chi - P^0 Sf \cdot T\chi) \, dx = -\int_\Omega f T\chi \, dx. \end{aligned} \tag{3.10}$$

Now (3.9) holds for all $\varphi \in H^1_P(\Omega)$ due to Lemma 3.3, in particular for $\varphi = Sf$; hence

$$\langle \gamma_1 Sf, \chi \rangle = \langle \psi, \gamma_0 Sf \rangle = 0.$$

This gives from (3.10)

$$\int_\Omega f T\chi \, dx = 0 \quad \text{for all } f \in L_2(\Omega).$$

Thus $T\chi = 0$ whence $\chi = \gamma_0 T\chi = 0$. From (3.9) now follows

$$\langle \psi, \gamma_0 \varphi \rangle = 0 \quad \text{for all } \varphi \in H^1(\Omega)$$

which implies $\psi = 0$ because of the surjectivity of

$$\gamma_0 : H^1(\Omega) \to H^{1/2}(\Gamma). \qquad \square$$

The continuity of the simple layer potential operator, Theorem 1(i) and (iii), will follow from an extension of Gagliardo's Trace Lemma, which seems to be new.

LEMMA 3.6. *For* $s \in (\frac{1}{2}, \frac{3}{2})$ *the trace map*

$$\gamma_0 : u \mapsto \gamma_0 u = u|_\Gamma : H^s_{\text{loc}}(\mathbb{R}^n) \to H^{s-1/2}(\Gamma) \quad \textit{is continuous.}$$

This result for $s = \frac{3}{2}$, from which the whole range $s \in (\frac{1}{2}, \frac{3}{2}]$ would follow by interpolation, is claimed by Jerison and Kenig [18]. However, there seems to be no proof available. The proof of Lemma 3.6 is given at the end of § 4.

The last tool we need is Nečas' result on the boundary regularity for the Dirichlet and Neumann problems.

LEMMA 3.7. *For* $\sigma \in [-\frac{1}{2}, \frac{1}{2}]$, *the mapping* $\gamma_1 T : H^{1/2+\sigma}(\Gamma) \to H^{-1/2+\sigma}(\Gamma)$ *is continuous, and* $\gamma_1 Tv \in H^{-1/2+\sigma}(\Gamma)$ *implies* $v \in H^{1/2+\sigma}(\Gamma)$.

Remarks. (i) Here the result for $\sigma < 0$ means, as above, the existence of a continuous extension of the map defined for $\sigma = 0$.

(ii) Nečas [23] showed that solutions of the Dirichlet problem with Dirichlet data in $H^1(\Gamma)$ have their Neumann data (i.e., conormal derivatives)in $L_2(\Gamma)$ and conversely. This is proved by applying an identity of Rellich, generalized to arbitrary second order equations by Payne and Weinberger. Thus it uses only partial integration and is completely elementary. The same argument has been used by Jerison and Kenig [16], [17] and Verchota [27], [21]. Having thus proved the result for $\sigma = \frac{1}{2}$, Nečas deduces the result for $\sigma = -\frac{1}{2}$ from a duality argument. The whole range $\sigma \in [-\frac{1}{2}, \frac{1}{2}]$ then clearly follows by interpolation.

4. The proofs.

Proof of Theorem 1 (i) *and* (iii). By definition (2.5) we can write the simple layer potential as

$$K_0 = G \circ \gamma_0', \tag{4.1}$$

where γ_0' is the adjoint of the trace map γ_0. By Lemma 3.6 we find that $\gamma_0 : H^{-s+1/2}(\Gamma) \to H^{-s}_{\text{comp}}(\mathbb{R}^n)$ is continuous for $s \in (\frac{1}{2}, \frac{3}{2})$. The operator G is a pseudodifferential operator of order -2 on $\mathbb{R}^n$, mapping $H^{-s}_{\text{comp}}(\mathbb{R}^n) \to H^{-s+2}_{\text{loc}}(\mathbb{R}^n)$ continuously for any $s \in \mathbb{R}$. Thus Theorem 1(i) follows. The continuity of the operator $A = \gamma_0 K_0$ then follows by a second application of Lemma 3.6. $\square$

Remark. If instead of Lemma 3.6, we use only the classical result (2.1), we find for theorem 1(i) only a range $\sigma \in [0, \frac{1}{2})$, and for (iii) only $\sigma = 0$ remains.

Next we use the representation formula (3.8) in order to write the double layer potential in terms of the simple layer potential. Writing (3.8) for a solution of the

Dirichlet problem $u = Tv \in H^1_P(\Omega)$ for $v \in H^{1/2}(\Gamma)$, we obtain $Tv = -K_1 v + K_0 \tilde{\gamma}_1 Tv$; hence

$$K_1 = (-1 + K_0 \widetilde{\gamma_1}) T. \tag{4.2}$$

This immediately implies that

$$K_1 : H^{1/2}(\Gamma) \to H^1(\Omega) \quad \text{is continuous.} \tag{4.3}$$

Thus, using (2.1) and Lemma 3.2, we obtain all statements of Theorem 1 for $\sigma = 0$ (i.e., in the "energy norm").

This will suffice to prove Theorem 2. The remaining cases of Theorem 1 will be shown together with Theorem 3.

Now we prove *jump relations* for the layer potentials. We use the notation introduced above

$$[\gamma_j u] := \gamma_j(u|_{\Omega^c}) - \gamma_j(u|_\Omega) \quad \text{for } j = 0, 1.$$

LEMMA 4.1.

$$[\gamma_0 K_0 \psi] = 0, \qquad [\gamma_1 K_0 \psi] = -\psi \quad \textit{for } \psi \in H^{-1/2}(\Gamma),$$

$$[\gamma_0 K_1 v] = v, \qquad [\tilde{\gamma}_1 K_1 v] = 0 \quad \textit{for } v \in H^{1/2}(\Gamma).$$

Proof. Let $\psi \in H^{-1/2}(\Gamma)$ and $u = K_0 \psi \in H^1_{\text{loc}}(\mathbb{R}^n)$. The equality $\gamma_0(u|_\Omega) = \gamma_0(u|_{\Omega^c})$ follows from the definition of γ_0. From (4.1) we obtain $Pu = \gamma_0' \psi$, if we apply P in the distributional sense to u. For any test function $\varphi \in C^\infty_{\text{comp}}(\mathbb{R}^n)$ we thus obtain

$$\int_{\mathbb{R}^n} uP'\varphi \, dx = \langle Pu, \varphi \rangle = \langle \gamma_0' \psi, \varphi \rangle = \langle \psi, \gamma_0 \varphi \rangle. \tag{4.4}$$

On the other hand, the second Green formula (3.7) gives

$$\int_\Omega uP'\varphi \, dx = \langle \widetilde{\gamma_1}(u|_\Omega), \gamma_0 \varphi \rangle - \langle \gamma_1 \varphi, \gamma_0 u \rangle.$$

The corresponding formula for Ω^c is

$$\int_{\Omega^c} uP'\varphi \, dx = -\langle \widetilde{\gamma_1}(u|_{\Omega^c}), \gamma_0 \varphi \rangle + \langle \gamma_1 \varphi, \gamma_0 u \rangle.$$

Adding both, we obtain with $[\gamma_0 u] = [\gamma_0 \varphi] = [\widetilde{\gamma_1} \varphi] = 0$

$$\int_{\mathbb{R}^n} uP'\varphi \, dx = -\langle [\widetilde{\gamma_1} u], \gamma_0 \varphi \rangle. \tag{4.5}$$

Comparison of (4.4) and (4.5) gives $[\gamma_1 u] = -\psi$, and from $[\gamma_0 u] = 0$ follows $[\gamma_1 u] = [\widetilde{\gamma_1} u] = -\psi$.

In order to show the jump relations for the double layer potential, we choose $v \in H^{1/2}(\Gamma)$ and $\varphi \in C^\infty_{\text{comp}}(\mathbb{R}^n)$ and define $u = K_1 v$. Then again the second Green formula gives

$$\int_{\mathbb{R}^n} uP'\varphi \, dx = -\langle [\widetilde{\gamma_1} u], \gamma_0 \varphi \rangle + \langle [\gamma_0 u], \gamma_1 \varphi \rangle. \tag{4.6}$$

On the other hand, the definition of K_1 gives $u = K_1 v = G(\gamma_1' v)$, where the compactly supported distribution on $\mathbb{R}^n$, $\gamma_1' v$ is defined by

$$\langle \gamma_1' v, \chi \rangle = \int_\Gamma v\, \partial_\nu \chi\, ds = \langle v, \gamma_1 \chi \rangle \quad \text{for all } \chi \in C^\infty_{\text{comp}}(\mathbb{R}^n).$$

Thus $Pu = \gamma_1' v$, implying

$$\int_{\mathbb{R}^n} uP'\varphi\, dx = \langle Pu, \varphi \rangle = \langle v, \gamma_1 \varphi \rangle. \tag{4.7}$$

Comparison of (4.6) and (4.7) gives

$$\langle v - [\gamma_0 u], \gamma_1 \varphi \rangle = \langle -[\tilde{\gamma}_1 u], \gamma_0 \varphi \rangle \quad \text{for all } \varphi \in C^\infty_{\text{comp}}(\mathbb{R}^n). \tag{4.8}$$

Finally we apply Lemma 3.5 which allows us to infer from (4.8):

$$v - [\gamma_0 u] = 0 = [\tilde{\gamma}_1 u]. \qquad \square$$

Proof of Theorem 2. Choose $v \in H^{-1/2}(\Gamma)$ and define $u = -K_0 v$. Then according to Lemma 4.1, we have the jump relations

$$[\gamma_0 u] = 0; \text{ hence } \gamma_0(u|_\Omega) = -Av = \gamma_0(u|_{\Omega^c}), \text{ and } [\gamma_1 u] = v. \tag{4.9}$$

Next we choose $\chi \in C^\infty_{\text{comp}}(\mathbb{R}^n)$ with $\chi = 1$ on a neighborhood of $\bar{\Omega}$ and define $u_1 := u|_\Omega$, $u_2 := \chi u|_{\Omega^c}$.

Next we add the first Green formula (3.6) for $u = v = u_1$ and its counterpart for Ω^c for $u = v = u_2$ and obtain using (4.9)

$$\Phi_\Omega(u_1, u_1) + \Phi_{\Omega^c}(u_2, u_2) - \int_{\Omega^c} \overline{u_2} P u_2\, dx = -\langle [\gamma_1 u], \overline{\gamma_0 u} \rangle = \langle v, \overline{Av} \rangle. \tag{4.10}$$

Here Φ_{Ω^c} is defined in accordance with (2.3).

Equation (4.10) now allows us to transfer the Gårding inequality for the operator P, which we assumed to hold, to the Gårding inequality on the boundary for the operator A.

The term $\int_{\Omega^c} \overline{u_2} P u_2\, dx$ gives rise to a compact bilinear form in $v \in H^{-1/2}(\Gamma)$ because Pu_2 has compact support in Ω^c and the mapping $v \mapsto u_2$ is continuous from $H^{-1/2}(\Gamma)$ to $C^\infty(\Omega^c)$. From the continuity of the trace mapping γ_1 (Lemma 3.2) we obtain an estimate

$$\begin{aligned} \|v\|^2_{H^{-1/2}(\Gamma)} &= \|\gamma_1 u_2 - \gamma_1 u_1\|^2_{H^{-1/2}(\Gamma)} \\ &\leqq C(\|u_1\|^2_{H^1(\Omega)} + \|u_2\|^2_{H^1(\Omega^c)} + \|Pu_1\|^2_{L_2(\Omega)} + \|Pu_2\|^2_{L_2(\Omega^c)}). \end{aligned} \tag{4.11}$$

Here on the right-hand side, Pu_1 vanishes and $\|Pu_2\|^2_{L_2(\Omega^c)}$ is a compact term.

Finally, the principal part of the right-hand side of (4.11) can be estimated from above up to compact terms by the left-hand side of (4.10) due to Gårding's inequality (2.4) which we assumed to hold. Thus (2.9) is proved.

In order to prove the strong ellipticity of the operator D, i.e., estimate (2.10), we proceed analogously.

For $v \in H^{-1/2}(\Gamma)$ we define $u = K_1 v$. Then we find the jump relations

$$[\gamma_0 u] = v \text{ and } [\overline{\gamma_1} u] = 0; \text{ hence } \overline{\gamma_1} u_1 = \overline{\gamma_1} u_2 = -Dv, \tag{4.12}$$

where u_1 and u_2 are defined from u as above. Then again the first Green formula gives

$$\Phi_\Omega(u_1, u_1) + \Phi_{\Omega^c}(u_2, u_2) - \int_{\Omega^c} \overline{u_2} P u_2\, dx = \langle Dv, \bar{v} \rangle. \tag{4.13}$$

This time the trace lemma (2.1) implies

$$\|v\|^2_{H^{1/2}(\Gamma)} = \|\gamma_0 u_2 - \gamma_0 u_1\|^2_{H^{1/2}(\Gamma)} \leqq C(\|u_1\|^2_{H^1(\Omega)} + \|u_2\|^2_{H^1(\Omega^c)}), \tag{4.14}$$

and again, (2.10) follows from (4.13) and (4.14) together with Gårding's inequality (2.4). □

The derivation of convergence and stability for Galerkin approximation schemes from strong ellipticity is standard by now [13], [26], as are the approximation properties of the finite element function spaces [1], [2]; thus proofs of Theorems 4 and 5 need not be given here.

Next we show regularity in the domain for the Dirichlet problem

LEMMA 4.2. *For $\sigma \in (-\frac{1}{2}, \frac{1}{2})$ the mapping $T: H^{1/2+\sigma}(\Gamma) \to H^{1+\sigma}_P(\Omega)$ is continuous.*

Proof. We choose a domain B containing $\bar{\Omega}$ in its interior, e.g., a large enough ball. Let $\Omega_2 := B \backslash \bar{\Omega}$ and $T_2: v \mapsto u = T_2 v$ be the solution operator of the Dirichlet problem

$$P^0 u = 0 \quad \text{in } \Omega_2, \quad \gamma_0 u = v, \quad u|_{\partial B} = 0.$$

Now choose $v \in H^1(\Gamma)$ and define

$$u = Tv \quad \text{in } \Omega, \qquad u = T_2 v \quad \text{in } \Omega_2.$$

Then the representation formula (3.8) applies and gives with $f = 0$ and $[\gamma_0 u] = 0$

$$u = -K_0[\widetilde{\gamma_1} u] + \int_{\partial B} \partial_\nu u(y) G(\cdot, y)\, ds(y) \quad \text{in } \Omega \cup \Omega_2. \tag{4.15}$$

Now we know from the boundary regularity result for the Dirichlet problem (Lemma 3.7) that there are estimates

$$\|\partial_\nu u|_{\partial B}\|_{H^{-1/2+\sigma}(\partial B)} + \|\widetilde{\gamma_1} Tv\|_{H^{-1/2+\sigma}(\Gamma)} + \|\widetilde{\gamma_1} T_2 v\|_{H^{-1/2+\sigma}(\Gamma)} \leqq C\|v\|_{H^{1/2+\sigma}(\Gamma)}$$

even for $\sigma \in [-\frac{1}{2}, \frac{1}{2}]$. Hence the continuity of the simple layer potential operator, Theorem 1(i) gives with (4.15) the desired estimate

$$\|u\|_{H^{1+\sigma}(\Omega)} \leqq C\|v\|_{H^{1/2+\sigma}(\Gamma)}. \qquad \square$$

Remark. The endpoint result $\sigma = \frac{1}{2}$ was shown by Jerison and Kenig [16] using Dahlberg's estimates for the Poisson kernel [10].

LEMMA 4.3. *For $s \in (\frac{1}{2}, \frac{3}{2})$ the trace map $\gamma_1: H^s_P(\Omega) \to H^{s-3/2}(\Gamma)$ is continuous.*

Proof. For $u \in H^1_P(\Omega)$ and arbitrary $\varphi \in H^{1/2}(\Gamma)$, $v := T\varphi$, we can apply the second Green formula (3.7) for the operator P^0 to obtain

$$\langle \gamma_1 u, \varphi \rangle = \langle \gamma_0 u, \gamma_1 T\varphi \rangle - \int_\Omega P^0 u T\varphi \, dx.$$

This can be written as

$$\gamma_1 = (\gamma_1 T)' \gamma_0 - T' P^0. \tag{4.16}$$

The first member on the right-hand side is continuous from $H^s(\Omega)$ to $H^{s-3/2}(\Gamma)$ due to Lemmas 3.6 and 3.7. The second member is continuous from $H^s(\Omega)$ to $H^t(\Gamma)$ for all s and $t < 0$ due to Lemma 4.2. □

As a corollary, Theorem 1(iv) follows from Theorem 1(i).

Proof of Theorem 1(ii), (v), and (vi). It suffices to show (ii). If we apply (4.2), this follows from Lemmas 4.2, 4.3, and Theorem 1(i). The proof of Theorem 1 is complete. □

Proof of Theorem 3. Let $\psi \in H^{-1/2}(\Gamma)$ and $A\psi \in H^{1/2+\sigma}(\Gamma)$. We show $\psi \in H^{-1/2+\sigma}(\Gamma)$. The a priori estimate (2.11) then follows from the closed graph theorem.

Define $u = K_0\psi$. Then we have $A\psi = \gamma_0 u$ and $\psi = -[\gamma_1 u]$ by Lemma 4.1. Thus u solves in Ω and Ω^c the Dirichlet problem with Dirichlet data $A\psi \in H^{1/2+\sigma}(\Gamma)$. According to Lemma 3.7, the Neumann data, and hence ψ, are in $H^{-1/2+\sigma}(\Gamma)$.

Now if $B\psi \in H^{-1/2+\sigma}(\Gamma)$, then $\gamma_1(u|_\Omega) = B\psi \in H^{-1/2+\sigma}(\Gamma)$, so that also $A\psi = \gamma_0(u|_\Omega) \in H^{1/2+\sigma}(\Gamma)$ holds.

The remaining statements follow in a similar way using the double layer potential and again Lemmas 3.7 and 4.1. □

Proof of Lemma 3.6. The statement is local, so we may assume that the boundary Γ is of the form

$$\Gamma = \{(x', x_n) \in \mathbb{R}^n \mid x' \in \mathbb{R}^{n-1};\ x_n = \psi(x')\}$$

with a function $\psi : \mathbb{R}^{n-1} \to \mathbb{R}$ that is uniformly Lipschitz, i.e., $\|\operatorname{grad} \psi\|_{L_\infty(\mathbb{R}^{n-1})} < \infty$.

For functions $f \in C^\infty_{\text{comp}}(\mathbb{R}^n)$ define

$$f_\psi(x', x_n) := f(x', x_n + \psi(x')).$$

We then have to show an estimate for $1 < s < \frac{3}{2}$

$$\|f_\psi(\cdot, 0)\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \leqq C\|f\|_{H^s(\mathbb{R}^n)} \quad \text{for all } f \in C^\infty_{\text{comp}}(\mathbb{R}^n). \tag{4.17}$$

The problem is that in general for $s > 1$, $f_\psi \notin H^s(\mathbb{R}^n)$ and therefore the usual trace lemma cannot be applied to f_ψ. We show first that the mapping $f \mapsto f_\psi$ leaves a certain anisotropic Sobolev space X^s invariant, and then that in X^s there holds the trace estimate

$$\|f(\cdot, 0)\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \leqq C\|f\|_{X^s} \quad \text{for all } f \in C^\infty_{\text{comp}}(\mathbb{R}^n). \tag{4.18}$$

For the definition of X^s we identify a function $f \in C^\infty_{\text{comp}}(\mathbb{R}^n)$ with the $C^\infty(\mathbb{R}^{n-1})$-valued function on $\mathbb{R}$,

$$x_n \mapsto f(\cdot, x_n).$$

Thus $H^s(\mathbb{R}^n) = H^s(\mathbb{R}; L_2(\mathbb{R}^{n-1})) \cap L_2(\mathbb{R}; H^s(\mathbb{R}^{n-1}))$. We define

$$X^s := H^s(\mathbb{R}; L_2(\mathbb{R}^{n-1})) \cap H^{s-1}(\mathbb{R}; H^1(\mathbb{R}^{n-1})).$$

If $\hat{f}$ is the Fourier transform of f, we define the norm in X^s by

$$\|f\|^2_{X^s} := \int_{-\infty}^{\infty} \int_{\mathbb{R}^{n-1}} \{(1+|\xi_n|)^{2s} + (1+|\xi_n|)^{2(s-1)} \cdot (1+|\xi'|)^2\} |\hat{f}(\xi', \xi_n)|^2 \, d\xi' \, d\xi_n.$$

Thus

$$\|f\|_{X^s} \leqq C\|f\|_{H^s(\mathbb{R}^n)}. \tag{4.19}$$

If we denote by $\tilde{f}(x', \xi_n)$ the Fourier transform of f with respect to the last variable, we have

$$\|f\|^2_{H^t(\mathbb{R};\, H^\tau(\mathbb{R}^{n-1}))} = \int_{-\infty}^{\infty} (1+|\xi_n|)^{2t} \|\tilde{f}(\cdot, \xi_n)\|^2_{H^\tau(\mathbb{R}^{n-1})} \, d\xi_n.$$

Now we have

$$\widetilde{(f_\psi)}(x', \xi_n) = e^{i\psi(x')\xi_n} \tilde{f}(x', \xi_n).$$

Hence

$$\|f_\psi\|_{H^t(\mathbb{R};\, L_2(\mathbb{R}^{n-1}))} = \|f\|_{H^t(\mathbb{R};\, L_2(\mathbb{R}^{n-1}))} \quad \text{for all } t \in \mathbb{R}. \tag{4.20}$$

For $k=1,\cdots,n-1$ we have

$$\widetilde{(\partial_k f_\psi)}(x',\xi_n)=e^{i\psi(x')\xi_n}\widetilde{(\partial_k f)}(x',\xi_n)+i\xi_n(\partial_k\psi)(x')\,e^{i\psi(x')\xi_n}\tilde f(x',\xi_n).$$

This implies that

$$\|\widetilde{(\partial_k f_\psi)}(\cdot,\xi_n)\|^2_{L_2(\mathbb{R}^{n-1})}\leqq\|\widetilde{(\partial_k f)}(\cdot,\xi_n)\|^2_{L_2(\mathbb{R}^{n-1})}+\xi_n^2\|\operatorname{grad}\psi\|^2_{L_\infty(\mathbb{R}^{n-1})}\|\tilde f(\cdot,\xi_n)\|^2_{L_2(\mathbb{R}^{n-1})}.$$

Hence

$$\|f_\psi\|^2_{H^t(\mathbb{R};\,H^1(\mathbb{R}^{n-1}))}\leqq\|f\|^2_{H^t(\mathbb{R};\,H^1(\mathbb{R}^{n-1}))}+C\|f\|^2_{H^{t+1}(\mathbb{R};\,L_2(\mathbb{R}^{n-1}))}\tag{4.21}$$

for all $t\in\mathbb{R}$.

Formulae (4.20) and (4.21) together imply the estimate

$$\|f_\psi\|_{X^s}\leqq\|f\|_{X^s}\quad\text{for all } s\in\mathbb{R}.\tag{4.22}$$

Next we show (4.18). We use the fact that with

$$m(\xi',\xi_n):=(1+|\xi_n|)^{2s}+(1+|\xi_n|)^{2s-2}\cdot(1+|\xi'|)^2$$

we have

$$\int_{-\infty}^{\infty}m(\xi',\xi_n)^{-1}\,d\xi_n=C_s(1+|\xi'|)^{1+2s}<\infty\quad\text{for } 1\leqq s<\tfrac{3}{2}.$$

Hence using the Cauchy-Schwarz inequality we have

$$\begin{aligned}\|f_\psi(\cdot,0)\|^2_{H^{s-1/2}(\mathbb{R}^{n-1})}&\int_{\mathbb{R}^{n-1}}(1+|\xi'|)^{2s-1}\left|\int_{-\infty}^{\infty}\hat f_\psi(\xi',\xi_n)\,d\xi_n\right|^2 d\xi'\\&\leqq\int_{\mathbb{R}^{n-1}}(1+|\xi'|)^{2s-1}\left(\int_{-\infty}^{\infty}m(\xi',\xi_n)^{-1}\,d\xi_n\right)\\&\quad\times\left(\int_{-\infty}^{\infty}m(\xi',\xi_n)|\hat f_\psi(\xi',\xi_n)|^2\,d\xi_n\right)d\xi'\\&=C\int_{\mathbb{R}^n}m(\xi',\xi_n)|\hat f_\psi(\xi',\xi_n)|^2\,d\xi_n\,d\xi'=C\|f_\psi\|^2_{X^s}.\end{aligned}$$

This together with the estimates (4.22) and (4.19) gives the desired estimate (4.17). □

5. Concluding remarks. (i) Along the same lines as presented here it is also possible to easily deduce invertibility results for integral equations involving the operators A, B, C, and D. Note that, for instance, by Theorem 2 the operators A and D are Fredholm operators of index 0 in the energy norm spaces. Thus if we assume injectivity, which in turn can be inferred from positivity of the bilinear form Φ_Ω, we obtain bijectivity. For the operator A this holds for the case of the Laplace equation and the standard fundamental solution in dimension $n\geqq 3$ and for $n=2$ if the analytic capacity of Γ is different from one. By Theorem 3 and duality arguments, bijectivity holds for the whole range of Sobolev spaces given in Theorem 1(ii). In this way we get results about the *solvability* of the boundary value problems by means of the boundary integral equations. Theorems 4 and 5 then are really statements about the numerical solution of boundary value problems by means of the so-called boundary element method [28]. In practice, this method is frequently used to solve (also mixed) boundary value problems of three-dimensional linear elasticity on domains with corners

and edges [25], [3]. For these problems, the present paper yields convergence proofs and asymptotic error estimates for Galerkin methods.

(ii) If the domain is more regular than merely Lipschitz, e.g., a smooth image of a polyhedron, then higher regularity results should be possible and they should improve with higher dimension. For the Dirichlet problem this is well known, but for the boundary integral equations higher regularity has been studied, to the best of the author's knowledge, only in the case of plane domains (see [6], [8], and the literature quoted therein).

REFERENCES

[1] J.-P. AUBIN, *Approximation of Elliptic Boundary Value Problems*, Wiley-Interscience, New York, 1972.

[2] A. K. AZIZ, ED., *The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.

[3] C. A. BREBBIA, J. C. F. TELLES, AND L. C. WROBEL, *Boundary Element Techniques*, Springer-Verlag, Berlin, 1984.

[4] J. CHAZARAIN AND A. PIRIOU, *Introduction à la théorie des equations aux dérivées partielles linéaires*, Gauthier-Villars, Paris, 1981.

[5] R. R. COIFMAN, A. MCINTOSH, AND I. MEYER, *L'intégrale de Cauchy définit un opérateur borné sur L^2 pour les courbes lipschitziennes*, Ann. of Math., 116 (1982), pp. 361-387.

[6] M. COSTABEL, *Boundary integral operators on curved polygons*, Ann. Mat. Pura Appl. (4), 133 (1983), pp. 305-326.

[7] ———, *Starke Elliptizität von Randintegraloperatoren erster Art*, Habilitationsschrift, THD-Preprint 982, Technische Hochschule, Darmstadt, West Germany, 1984.

[8] M. COSTABEL, E. STEPHAN, AND W. L. WENDLAND, *On boundary integral equations of the first kind for the bi-Laplacian in a polygonal plane domain*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 10 (1983), pp. 197-241.

[9] M. COSTABEL AND W. L. WENDLAND, *Strong ellipticity of boundary integral operators*, J. Reine Angew. Math., 372 (1986), pp. 39-63.

[10] B. E. J. DAHLBERG, *On the Poisson integral for Lipschitz and C^1 domains*, Studia Math., 66 (1979), pp. 7-24.

[11] J. DIEUDONNÉ, *Eléments d'analyse*, Vol. 8, Gauthier-Villars, Paris, 1978.

[12] E. B. FABES, M. JODEIT, AND N. M. RIVIERE, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165-186.

[13] I. C. GOHBERG AND J. A. FELDMAN, *Convolution Equations and Projection Methods for their Solution*, American Mathematical Society, Providence, RI, 1974.

[14] P. GRISVARD, *Boundary Value Problems in Non-Smooth Domains*, Pitman, London, 1985.

[15] G. C. HSIAO AND W. L. WENDLAND, *A finite element method for some integral equations of the first kind*, J. Math. Anal. Appl., 58 (1977), pp. 449-481.

[16] D. S. JERISON AND C. E. KENIG, *The Dirichlet problem in nonsmooth domains*, Ann. of Math., 113 (1981), pp. 367-382.

[17] ———, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc., 4 (1981), pp. 203-207.

[18] ———, *Boundary value problems on Lipschitz domains*, in Studies in Partial Differential Equations, MAA Studies in Mathematics 23, W. Littmann, ed., Math. Assoc. of America, 1982, pp. 1-68.

[19] J. L. JOURNÉ, *Calderón-Zygmund Operators, Pseudo-Differential Operators and the Cauchy Integral of Calderón*, Lecture Notes in Mathematics 994, Springer-Verlag, Berlin, 1983.

[20] C. E. KENIG, *Boundary value problems of linear elastostatics and hydrostatics on Lipschitz domains*, Sem. Goulaouic-Meyer-Schwartz 1983-1984, Exposé n° XXI.

[21] I. MEYER, *Théorie du potentiel dans les domaines Lipschitziens d'après G. C. Verchota*, Sem. Goulaouic-Meyer-Schwartz 1982-1983, Exposé n° V.

[22] J. C. NEDELEC, *Approximation des equations integrales en méchanique et physique*, Rapport Interne, Ecole Polytechnique, Palaiseau, 1977.

[23] J. NEČAS, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[24] L. E. PAYNE AND H. F. WEINBERGER, *New bounds for solutions of second order elliptic partial differential equations*, Pacific J. Math., 8 (1958), pp. 551-573.

[25] P. C. RIZZONELLI, *On the first boundary value problem for the classical theory of elasticity in a three-dimensional domain with a singular boundary*, J. Elasticity, 3 (1973), pp. 225-259.

[26] E. STEPHAN AND W. L. WENDLAND, *Remarks to Galerkin and least squares methods with finite elements for general elliptic problems*, Manuscripta Geodaetica, 1 (1976), pp. 93–123.

[27] G. VERCHOTA, *Layer potentials and regularity for the Dirichlet problem for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

[28] W. L. WENDLAND, *Boundary element methods and their asymptotic convergence*, in Theoretical Acoustics and Numerical Techniques, CISM Courses 277, P. Filippi, ed., Springer-Verlag, Vienna, New York, 1983, pp. 135–216.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

009

ON THE SHARPNESS OF WEYL'S ESTIMATES FOR EIGENVALUES OF SMOOTH KERNELS, II*

J. B. READE†

Abstract. The estimate $\lambda_n = o(1/n)$ obtained by H. Weyl (1912) for the nth largest in modulus eigenvalue λ_n of any symmetric Fredholm operator on $L^2[0, 1]^2$ with kernel in $C^1[0, 1]^4$ is shown to be best possible in the sense that for any increasing sequence $\alpha_n \to \infty$ there exist such operators whose nth eigenvalue is not $o(1/n\alpha_n)$. The construction of the counterexample makes use of Rudin-Shapiro polynomials. The corresponding result for positive definite operators is proved with a simpler counterexample. The methods generalise to the case $L^2[0, 1]^m (m \geqq 3)$ without further difficulty.

Key words. eigenvalue, operator, kernel, asymptotics

AMS(MOS) subject classification. 45C

1. Introduction.

If $K(x, t) = \overline{K(t, x)} \in L^2[a, b]^2$ then

$$(Tf)(x) = \int_a^b K(x, t) f(t)\, dt$$

defines a compact symmetric operator T on the Hilbert space $L^2[a, b]$. Such an operator T has a real null sequence $(\lambda_n)_{n \geqq 1}$ of eigenvalues which we can assume has been enumerated so that

$$|\lambda_1| \geqq |\lambda_2| \geqq \cdots \geqq |\lambda_n| \geqq \cdots .$$

H. Weyl showed in [4] that if $K(x, t) \in C^1[a, b]^2$, i.e., $K(x, t)$ has continuous partial derivatives, then $\lambda_n = o(1/n^{3/2})$. We showed in [1] that if $K(x, t) \in PDC^1[a, b]^2$, i.e., $K(x, t)$ is positive definite and $\in C^1[a, b]^2$, then $\lambda_n = o(1/n^2)$.

Similar results are true for operators of the form

$$(Tf)(x, y) = \int_a^b \int_a^b K(x, y, t, u) f(t, u)\, dt\, du$$

where $K(x, y, t, u) = \overline{K(t, u, x, y)} \in L^2[a, b]^4$ and $f(t, u) \in L^2[a, b]^2$. The estimates are $\lambda_n = o(1/n)$ for $K(x, y, t, u) \in C^1[a, b]^4$ and $\lambda_n = o(1/n^{3/2})$ for $K(x, y, t, u) \in PDC^1[a, b]^4$. The proofs are similar to those for kernels in two variables.

In [2] we considered the sharpness of the estimates for two variable kernels. Here we consider four variable kernels.

2. Double Fourier series.

Any $k(t, u) \in L^2[0, 1]^2$ has a double Fourier series

$$\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} c_{mn}\, e^{2\pi i(mt+nu)}$$

where

$$c_{mn} = \int_0^1 \int_0^1 k(t, u)\, e^{-2\pi i(mt+nu)}\, dt\, du$$

* Received by the editors February 25, 1986; accepted for publication (in revised form) May 12, 1987.
† Department of Mathematics, The University of Manchester, Manchester, United Kingdom M13 9PL.

for all integers m, n. The series is unconditionally convergent to $k(t, u)$ in mean square, so in particular

$$k(t, u) = \lim_{N\to\infty} \sum_{m=-N}^{N} \sum_{n=-N}^{N} c_{mn}\, e^{2\pi i(mt+nu)}$$

(in mean square).

LEMMA 1. *If $k(t, u) \in L^2[0, 1]^2$ has double Fourier series*

$$k(t, u) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} c_{mn}\, e^{2\pi i(mt+nu)}$$

then $K(x, y, t, u) = k(x-t, y-u)$ has eigenvalues c_{mn} and eigenfunctions $e^{2\pi i(mt+nu)}$ (m, n integers).

Proof. For any integers p, q

$$\begin{aligned}
&\int_0^1 \int_0^1 K(x, y, t, u)\, e^{2\pi i(pt+qu)}\, dt\, du \\
&\quad = \lim_{N\to\infty} \int_0^1 \int_0^1 \sum_{m=-N}^{N} \sum_{n=-N}^{N} c_{mn}\, e^{2\pi im(x-t)}\, e^{2\pi in(y-u)}\, e^{2\pi i(pt+qu)}\, dt\, du \\
&\quad = c_{pq}\, e^{2\pi i(px+qy)}.
\end{aligned}$$

Observe that $K(x, y, t, u)$ is symmetric if $k(t, u) = \overline{k(-t, -u)}$, equivalently if c_{mn} is real for all m, n.

3. The Rudin–Shapiro signs. These are $\varepsilon_n = \pm 1 (n \geqq 0)$ with the property that

$$\sum_{n=0}^{N} \varepsilon_n\, e^{2\pi int} = O(N^{1/2})$$

uniformly in t. They occur as the coefficients of successive polynomials $P_n(z)$ defined inductively with $Q_n(z)$ by saying

$$\begin{aligned}
&P_0(z) = Q_o(z) = 1, \\
&P_{n+1}(z) = P_n(z) + z^{2^n} Q_n(z), \\
&Q_{n+1}(z) = P_n(z) - z^{2^n} Q_n(z) \qquad (n \geqq 1).
\end{aligned}$$

(See [3] for details.)

LEMMA 2. *If a_{mn} is the m, nth entry of the matrix*

$$\begin{pmatrix}
1 & 1/2^\alpha & 1/3^\alpha & \cdots & 1/n^\alpha & \cdots \\
1/2^\alpha & 1/2^\alpha & 1/3^\alpha & & 1/n^\alpha & \\
1/3^\alpha & 1/3^\alpha & 1/3^\alpha & & 1/n^\alpha & \\
\vdots & & & & \vdots & \\
1/n^\alpha & 1/n^\alpha & 1/n^\alpha & \cdots & 1/n^\alpha & \cdots \\
\vdots & & & & \vdots &
\end{pmatrix}$$

then the double Fourier series

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \varepsilon_m \varepsilon_n a_{mn}\, e^{2\pi i(mt+nu)}$$

is uniformly convergent in t, u for all $\alpha > 1$.

Proof. Let

$$A_N(t, u) = \sum_{m=1}^{N} \sum_{n=1}^{N} \varepsilon_m \varepsilon_n a_{mn}\, e^{2\pi i(mt+nu)}$$

for each $N \geqq 1$. Then if we write

$$s_N(t) = \sum_{n=1}^{N} \varepsilon_n\, e^{2\pi int}$$

we have

$$A_N(t, u) - A_M(t, u) = \sum_{n=M+1}^{N} \frac{s_n(t)s_n(u) - s_{n-1}(t)s_{n-1}(u)}{n^\alpha}$$

$$= -\frac{s_M(t)s_M(u)}{(M+1)^\alpha} + \sum_{n=M+1}^{N-1} \left(\frac{1}{n^\alpha} - \frac{1}{(n+1)^\alpha}\right) s_n(t)s_n(u) + \frac{s_N(t)s_N(u)}{N^\alpha}.$$

The first and third terms are $O(1/M^{\alpha-1})$, $O(1/N^{\alpha-1})$, respectively, and the middle term is the difference between the $(N-1)$th and Mth partial sums of a series whose nth term is $O(1/n^\alpha)$, all uniformly in t, u.

COROLLARY. *The function*

$$k(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \varepsilon_m \varepsilon_n a_{mn}\, e^{2\pi i(mt+nu)},$$

where a_{mn} is as in Lemma 2, is continuous if $\alpha > 1$.

LEMMA 3. *The derivative*

$$s'_N(t) = \sum_{n=1}^{N} 2\pi in\varepsilon_n\, e^{2\pi int} = O(N^{3/2})$$

uniformly in t.

Proof. We have

$$s'_N(t) = \sum_{1}^{N} 2\pi in(s_n(t) - s_{n-1}(t))$$

$$= -2\pi i \left(\sum_{1}^{N} s_n(t)\right) + 2\pi iNs_N(t)$$

$$= \left(\sum_{1}^{N} O(n^{1/2})\right) + NO(N^{1/2})$$

$$= O(N^{3/2}).$$

COROLLARY. *The function*

$$k(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \varepsilon_m \varepsilon_n a_{mn}\, e^{2\pi i(mt+nu)},$$

where a_{mn} is as in Lemma 2, is C^1 if $\alpha > 2$.

Proof. If

$$B_N(t, u) = \sum_{m=1}^{N} \sum_{n=1}^{N} 2\pi im\varepsilon_m \varepsilon_n a_{mn}\, e^{2\pi i(mt+nu)}$$

is the Nth partial sum of the series obtained by formally differentiating the double Fourier series of $k(x, t)$ partially with respect to t, then we have

$$B_N(t, u) - B_M(t, u) = \sum_{n=M+1}^{N} \frac{s_n'(t)s_n(u) - s_{n-1}'(t)s_{n-1}(u)}{n^\alpha}$$

$$\to 0$$

as $M, N \to \infty$ uniformly in t, u by the same argument as that used in the proof of Lemma 2. It follows that

$$\frac{\partial k}{\partial t}(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} 2\pi i m \varepsilon_m \varepsilon_n a_{mn} e^{2\pi i(mt+nu)}$$

is continuous, and similarly so is the other partial derivative.

LEMMA 4. *For any given $\alpha > 1$ there exist C^1 kernels $K(x, y, t, u)$ whose eigenvalues are not $o(1/n^\alpha)$.*

Proof. If $\alpha > 2$ and

$$k(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \varepsilon_m \varepsilon_n a_{mn} e^{2\pi i(mt+nu)},$$

where a_{mn} is as in Lemma 2, then $K(x, y, t, u) = k(x-t, y-u)$ is C^1 and has eigenvalues $\varepsilon_m \varepsilon_n a_{mn} (m, n \geqq 1)$. Arranging these eigenvalues in descending order of modulus and denoting them by

$$|\lambda_1| \geqq |\lambda_2| \geqq \cdots \geqq |\lambda_n| \geqq \cdots$$

we have $|\lambda_{n^2}| = 1/n^\alpha$ and so λ_n is not $o(1/n^{\alpha/2})$.

4. Positive definite kernels.

LEMMA 5. *If b_{mn} is the m, nth entry of the matrix*

$$\begin{pmatrix} 1 & 1/2^\alpha & 1/3^\alpha & \cdots & 1/n^\alpha & \cdots \\ 1/2^\alpha & 1/3^\alpha & & & & \\ 1/3^\alpha & & & & & \\ \vdots & & & & & \\ 1/n^\alpha & & & & & \\ \vdots & & & & & \end{pmatrix}$$

then the double Fourier series

$$\sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn} e^{2\pi i(mt+nu)}$$

is uniformly absolutely convergent in t, u for all $\alpha > 2$.

Proof. The sum of the entries b_{mn} on the nth cross-diagonal of the matrix is $n/n^\alpha = 1/n^{\alpha-1}$ and $\sum_1^\infty 1/n^{\alpha-1} < \infty$ if $\alpha > 2$.

COROLLARY. *The function*

$$k(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn} e^{2\pi i(mt+nu)},$$

where b_{mn} is as in Lemma 5, is continuous if $\alpha > 2$.

LEMMA 6. *The function*

$$k(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn}\, e^{2\pi i(mt+nu)},$$

where b_{mn} is as in Lemma 5, is C^1 if $\alpha > 3$.

Proof. Formally we have

$$\frac{\partial k}{\partial t}(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} 2\pi i m b_{mn}\, e^{2\pi i(mt+nu)}$$

and the sum of the entries on the nth cross-diagonal is now

$$2\pi i \frac{1+2+\cdots+n}{n^{\alpha}} = O(1/n^{\alpha-2})$$

and $\sum_1^{\infty} 1/n^{\alpha-2} < \infty$ if $\alpha > 3$.

LEMMA 7. *For any given $\alpha > 3/2$ there exist PDC^1 kernels $K(x, y, t, u)$ whose eigenvalues are not $o(1/n^{\alpha})$.*

Proof. If $\alpha > 3$ and

$$k(t, u) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} b_{mn}\, e^{2\pi i(mt+nu)},$$

where b_{mn} is as in Lemma 5, then $K(x, y, t, u) = k(x-t, y-u)$ is PDC^1 and has eigenvalues $b_{mn}(m, n \geqq 1)$. Arranging in descending order and denoting by

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n \geqq \cdots$$

we have $\lambda_{n(n+1)/2} = 1/n^{\alpha}$ and so λ_n is not $o(1/n^{\alpha/2})$.

5. Sharper results.

LEMMA 8. *For any given real sequence $(\alpha_n)_{n\geqq 1}$ which increases and diverges to infinity there exist C^1 kernels $K(x, y, t, u)$ whose eigenvalues are not $o(1/n\alpha_n)$.*

Proof (see [2]). Choose $n_k\ (k \geqq 1)$ such that $\alpha_{n_k} > k^2$ and let $\beta_n = 1/n_k^2 \alpha_{n_k} (n_{k-1} < n \leqq n_k)$. If we replace every entry in the nth L-shaped section of the matrix (a_{mn}) of Lemma 2 by β_n, then the corresponding kernel $K(x, y, t, u)$ constructed as in Lemma 4 has the required properties.

LEMMA 9. *For any given increasing divergent real sequence $(\alpha_n)_{n\geqq 1}$ there exist PDC^1 kernels $K(x, y, t, u)$ whose eigenvalues are not $o(1/n^{3/2}\alpha_n)$.*

Proof. Choose $n_k (k \geqq 1)$ as in the proof of Lemma 8, but now let $\beta_n = 1/n_k^3 \alpha_{n_k} (n_{k-1} < n \leqq n_k)$ and replace every entry in the nth cross-diagonal of the matrix (b_{mn}) of Lemma 5 by β_n.

Acknowledgments. The author would like to thank Graham Little and Fritz Ursell for useful conversations, and the referee for pointing out an error in § 3.

REFERENCES

[1] J. B. READE, *Eigenvalues of positive definite kernels*, this Journal, 14 (1983), pp. 152–157.

[2] ———, *On the sharpness of Weyl's estimate for eigenvalues of smooth kernels*, this Journal, 16 (1985), pp. 548–550.

[3] W. RUDIN, *Some theorems on Fourier coefficients*, Proc. Amer. Math. Soc., 10 (1959), pp. 855–859.

[4] H. WEYL, *Das Asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen*, Math. Ann., 71 (1912), pp. 441–479.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

010

ON OPTIMAL ALGORITHMS IN AN ASYMPTOTIC MODEL WITH GAUSSIAN MEASURE*

G. W. WASILKOWSKI† AND H. WOŹNIAKOWSKI‡

Abstract. We study the approximate solution of linear problems in separable Hilbert spaces equipped with a Gaussian measure. We find information and algorithms with the best possible rate of convergence. Although adaptive information and nonlinear algorithms are permitted, we prove that *nonadaptive* information and *linear* algorithms are optimal. An algorithm is optimal if it converges with a rate of convergence that is no worse than the rate of any other algorithm except on sets of measure zero. We prove that algorithms and information that minimize the average errors lead to the best possible rate of convergence. This exhibits a close relation between the asymptotic and average case models.

Key words. asymptotic model, linear problems, Gaussian measures

AMS(MOS) subject classifications. 65J05, 41A65

1. Introduction. Many papers deal with optimal algorithms for problems that are approximately solved. In these papers an optimal algorithm is usually defined as one having minimal error. In a *worse case model* the error of an algorithm is defined by its worst performance whereas in an *average case model* the error of an algorithm is defined by its average performance (see [7]-[11]).

To make clear what we mean by worst and average case models we consider a simple integration example.

Example 1.1. Suppose we want to approximate $Sf = \int_0^1 f(t)\,dt$ where $f:[0,1] \to \mathbf{R}$ is a function belonging to a given class F, where F is a subset of a Hilbert space F_1. We assume that we sample f at n given points $t_1, t_2, \cdots, t_n$. Thus we know that $N_n(f) = [f(t_1), f(t_2), \cdots, f(t_n)]$. Based on $N_n(f)$ and the fact that $f \in F$, we approximate $S(f)$ by an algorithm φ_n. By an algorithm we mean a mapping from $N_n(F)$ into $\mathbf{R}$. Thus, $\varphi_n(N_n(f)) = \sum_{i=1}^n a_i f(t_i)$ is an example of an algorithm. In the worst case model, the error of φ_n is defined as:

$$e^w(\varphi_n, N_n) = \sup\{|Sf - \varphi_n(N_n(f))| : f \in F\}.$$

In the average case model, the error of φ_n is usually defined as:

$$e^{\text{avg}}(\varphi_n, N_n) = \sqrt{\int_F |Sf - \varphi_n(N_n(f))|^2 \mu(df)}$$

where μ is a given probability measure on F.

In this paper we study optimal algorithms in the asymptotic model. To explain what we mean by optimality in the asymptotic model, we shall use the integration example.

We first stress the main difference between the asymptotic and worst (or average) case models. In the worst (or average) case model fixed information N_n is applied for all functions f from F. This may be contrasted with the asymptotic model, where a

* Received by the editors December 1, 1986; accepted for publication (in revised form) May 21, 1987. This research was supported in part by the National Science Foundation under grants DCR-86-03674 and ICT-85-17289.

† Department of Computer Science, University of Kentucky, Lexington, Kentucky 40506.

‡ Institute of Informatics, University of Warsaw, Warsaw, Poland, and Department of Computer Science, Columbia University, New York, New York 10027.

sequence of information N_n, with n tending to infinity, is applied for each f. That is, knowing $N_n(f)$ for all n, we want to find a sequence of approximations $\varphi_n(N_n(f))$ to Sf with good convergence properties. Here φ_n is an algorithm that uses N_n. Let $\bar{\varphi} = \{\varphi_n\}$ be a sequence of such algorithms. For brevity, $\bar{\varphi}$ is also called an algorithm.

Asymptotically convergent algorithms are widely used in practice. Examples include quadrature formulas for the integration problem, and algorithms for the solution of ordinary or partial differential equations with meshsize tending to zero. This approach, commonly used in numerical analysis, motivates our study. Our interest is to find an asymptotically optimal algorithm.

What is the optimal algorithm in the asymptotic model? We motivate our definition by the following discussion. Let $\bar{\varphi} = \{\varphi_n\}$ and $\bar{\varphi}^* = \{\varphi_n^*\}$ be two algorithms for approximating Sf. Let $A(\bar{\varphi}, \bar{\varphi}^*)$ denote the set of functions f for which the algorithm $\bar{\varphi}$ converges to Sf with a better rate of convergence than the algorithm $\bar{\varphi}^*$. That is, $f \in A(\bar{\varphi}, \bar{\varphi}^*)$ if and only if

$$\lim_n |Sf - \varphi_n(N_n(f))| / |Sf - \varphi_n^*(N_n(f))| = 0.$$

One might want to define $\bar{\varphi}^*$ as optimal if $\bar{\varphi}^*$ *never* "loses" to $\bar{\varphi}$. That is, $\bar{\varphi}^*$ is optimal if $A(\bar{\varphi}, \bar{\varphi}^*) = \varnothing$, for all $\bar{\varphi}$. We show that such an algorithm $\bar{\varphi}^*$ does *not* exist. Another attempt would be to define $\bar{\varphi}^*$ as optimal by requiring that $A(\bar{\varphi}, \bar{\varphi}^*)$ be finite or perhaps countable for all $\bar{\varphi}$. Unfortunately this also does not work. (This is proven in the Appendix.) Thus, optimality of $\bar{\varphi}^*$ must be defined differently.

One approach that does work is due to Trojan [8]. He defines $\bar{\varphi}^*$ to be optimal if the set $A(\bar{\varphi}, \bar{\varphi}^*)$ has empty interior for any $\bar{\varphi}$. We define optimality differently than Trojan. We assume that the space of elements f is equipped with a probability measure μ. Optimality of $\bar{\varphi}^*$ is then defined by zero measure of the set $A(\bar{\varphi}, \bar{\varphi}^*)$ for any algorithm $\bar{\varphi}$. Thus, $\bar{\varphi}^*$ is optimal if it "loses" to any algorithm $\bar{\varphi}$ only on a set of measure zero.

The integration example discussed above is a particular case of problems studied in this paper. The general formulation is as follows. For two separable Hilbert spaces F_1 and F_2, we consider a linear continuous operator S, $S: F_1 \to F_2$. We wish to approximate Sf for any element f from F_1. We assume that the element f is not known. Instead one can compute n linear continuous functionals $N_n(f)$. The choice of the ith functional may depend on the values of $(i-1)$ previously computed functionals. Then the sequence $\bar{N} = \{N_n\}$ is called *adaptive* information. For given adaptive information $\bar{N}$ we wish to find a sequence $\bar{\varphi} = \{\varphi_n\}$ such that $\varphi_n(N_n(f))$ goes to Sf with the best possible rate of convergence. The sequence $\bar{\varphi}$ is called an *algorithm*. We are looking for an *optimal* algorithm $\bar{\varphi}^* = \{\varphi_n^*\}$. As we already mentioned, optimality of $\bar{\varphi}^*$ means that an arbitrary algorithm $\bar{\varphi}$ approximates Sf with a better rate of convergence than the algorithm $\bar{\varphi}^*$ only on a set of measure zero. More precisely, we assume that the space F_1 is equipped with a Gaussian measure μ. Then $\bar{\varphi}^*$ is (asymptotically) *optimal* if and only if

$$\mu(\{f \in F_1 : \lim_n \|Sf - \varphi_n(N_n(f))\| / \|Sf - \varphi_n^*(N_n(f))\| = 0\}) = 0 \tag{1.1}$$

for any algorithm $\bar{\varphi} = \{\varphi_n\}$. Here we adopt the convention $0/0 = 1$.

Obviously, an optimal algorithm is *not* uniquely defined by (1.1). However, the difference between optimal algorithms is insignificant. Indeed, let $h_n = \varphi_{1,n} - \varphi_{2,n}$ be the difference between two optimal algorithms $\bar{\varphi}_1 = \{\varphi_{1,n}\}$ and $\bar{\varphi}_2 = \{\varphi_{2,n}\}$. Then $\|h_n(N_n(f))\| \leqq \|Sf - \varphi_{1,n}(N_n(f))\| + \|Sf - \varphi_{2,n}(N_n(f))\|$ goes to zero at least as fast as the errors of optimal algorithms.

The first problem studied in this paper is to find an optimal algorithm in the sense of (1.1). We solve this problem by showing a relation between optimality in the average case and asymptotic models. More precisely, let φ_n^* be an optimal algorithm in the average case model. The form of φ_n^* is known (see [10], [11]). In fact, $\varphi_n^*(N_n(f)) = S\sigma_n(N_n(f))$ where $\sigma_n(N_n(f))$ is a μ-spline element. If information N_n is nonadaptive, then φ_n^* is a linear mapping.

Let $\bar{\varphi}^* = \{\varphi_n^*\}$. That is, the algorithm $\bar{\varphi}^*$ consists of algorithms φ_n^* that minimize the average case errors for each n. The algorithm $\bar{\varphi}^*$ is called a μ-spline algorithm since it is based on μ-spline elements. We prove that $\bar{\varphi}^*$ is optimal. This is very desirable from a practical point of view. The μ-spline algorithm that has the best possible rate of convergence also has the minimal average case error at each step.

The second problem studied in this paper is to characterize the rate of convergence of the μ-spline algorithm. Once more, we solve this problem by showing a relation to the average case model. This relation is exhibited in terms of local average radii which play a key role in the average case analysis. We show that the sequence of local average radii fully characterizes the rate of convergence of the μ-spline algorithm.

The third problem is to find information $\bar{N}^* = \{N_n^*\}$ for which the rate of convergence of local radii is best possible, or equivalently, for which the rate of convergence of the optimal algorithm is best possible. It turns out that N_n^* is given by just that *nonadaptive* information that is optimal in the average case model.

Thus, although we permit adaptive information and nonlinear algorithms, the optimal information is *nonadaptive* and the optimal algorithm is *linear.*

Our results exhibit a close relation between the average case model and the asymptotic model for linear problems defined in Hilbert spaces with a Gaussian measure. Trojan [8] shows a similar relation between the worst case and the asymptotic models for linear problems defined on a Banach space that is *not* equipped with a measure. As we already mentioned, in his paper asymptotic optimality is defined by the condition that the sets $A(\bar{\varphi}, \bar{\varphi}^*)$ have empty interior. These relations are desirable from a practical point of view. Algorithms and information that minimize the worst or average case error also yield the best rate of convergence.

Our results hold for linear problems defined in Hilbert spaces with a Gaussian measure. We shall indicate which of them hold for more general measures, which are called elliptically contoured (see [1]). Recent research seems to indicate that the results of the paper also hold if F_1 is a Banach space.

The contents of this paper are summarized as follows. In § 2 we formulate the three problems studied in this paper. In the successive sections we present solutions to these problems. In § 3 we prove optimality of the μ-spline algorithm. In § 4 we show that the local average radii fully characterize the rate of convergence of the μ-spline algorithm. In § 5 we exhibit optimal information. Finally, the Appendix contains a proof that shows that the set $A(\bar{\varphi}, \bar{\varphi}^*)$ is uncountable.

2. Formulation of the problem. In this section we define the basic concepts and formulate the three problems that will be studied in this paper.

Let F_1 and F_2 be separable Hilbert spaces over the real field. Let μ be a Gaussian measure defined on Borel sets of F_1, $\mu(F_1) = 1$. We assume that the mean element of μ is zero and the self-adjoint covariance operator S_μ of μ is positive definite (see [3], [6]).

Consider a continuous linear operator S such that

$$S: F_1 \to F_2. \tag{2.1}$$

Our problem is to approximate Sf for all f from F_1. We assume that the element f is

not known but we can compute arbitrarily many functionals of f. That is, we define an *information operator* $\bar{N}$ of the form

$$\bar{N}(f) = [(f, g_1), (f, g_2(y_1)), \cdots, (f, g_n(y_1, \cdots, y_{n-1})), \cdots] \tag{2.2}$$

where $y_1 = (f, g_1)$, $y_i = (f, g_i(y_1, \cdots, y_{i-1}))$ for $i = 2, 3, \cdots$. Here $(\cdot, \cdot)$ denotes the inner product in F_1 and $g_i: \mathbf{R}^{i-1} \to F_1$ is a measurable mapping, i.e., $g_i^{-1}(B)$ is a Borel set of $\mathbf{R}^{i-1}$ whenever B is a Borel set of F_1. Without loss of generality (see [10], [11]) we assume that

$$(S_\mu g_i(y_1, \cdots, y_{i-1}), g_j(y_1, \cdots, y_{j-1})) = \delta_{i,j} \tag{2.3}$$

for $i, j = 1, 2, \cdots$, and all $y_1, y_2, \cdots$.

The essence of (2.2) is that the choice of $g_i(y_1, \ldots, y_{i-1})$ may depend on the $i-1$ previously computed inner products. Such an information operator $\bar{N}$ is therefore called *adaptive*. To stress the adaptive character of $\bar{N}$ we shall sometimes write $\bar{N} = \bar{N}^a$. On the other hand, if each $g_i(y_1, \cdots, y_{i-1})$ does not depend on $y_1, \cdots, y_{i-1}$, i.e., $g_i(y_1, \cdots, y_{i-1}) \equiv g_i^*$ for some g_i^* from F_1, then $\bar{N}$ is called *nonadaptive* and sometimes denoted by $\bar{N} = \bar{N}^{\text{non}}$. Thus the information operator $\bar{N}$ consists of a sequence of inner products chosen adaptively or nonadaptively. In either case, let

$$N_n(f) = [(f, g_1), (f, g_2(y_1)), \cdots, (f, g_n(y_1, \cdots, y_{n-1}))] \quad \forall f = F_1, \tag{2.4}$$

denote the first n inner products in $\bar{N}$. Knowing $N_n(f)$ for all n, we approximate Sf by an algorithm $\bar{\varphi}$. By the *algorithm* $\bar{\varphi} = \{\varphi_n\}$ that uses $\bar{N} = \{N_n\}$ we mean a sequence of mappings

$$\varphi_n: N_n(F_1) \subset \mathbf{R}^n \to F_2. \tag{2.5}$$

We assume that φ_n is measurable, i.e., $\varphi_n^{-1}(B)$ is a Borel set of $\mathbf{R}^n$ whenever B is a Borel set of F_2.

We approximate Sf by $\varphi_n(N_n(f))$. We want to find $\bar{N}$ and $\bar{\varphi}$ such that $\varphi_n(N_n(f))$ tends to Sf as fast as possible. More precisely, we shall study the following three problems:

(i) For a given information operator $\bar{N} = \{N_n\}$ find an *optimal algorithm* $\bar{\varphi}^* = \{\varphi_n^*\}$ using $\bar{N} = \{N_n\}$. That is, an algorithm $\bar{\varphi}^*$ such that

$$\mu\left(\left\{f \in F_1: \lim_n \frac{\|Sf - \varphi_n(N_n(f))\|}{\|Sf - \varphi_n^*(N_n(f))\|} = 0\right\}\right) = 0 \tag{2.6}$$

for any algorithm $\bar{\varphi} = \{\varphi_n\}$ using $\bar{N}$. Throughout this paper we adopt the convention $0/0 = 1$.

(ii) Let $\bar{\varphi}^*$ satisfy (2.6). Characterize the rate of convergence of φ^* in terms of the Gaussian measure μ, the operator S and the information operator $\bar{N}$.

(iii) Find an *optimal* information operator $\bar{N}^*$, i.e., $\bar{N}^*$ of the form (2.2) for which the rate of convergence of an optimal algorithm $\bar{\varphi}^*$ using $\bar{N}^*$ is best possible. This means, we want to determine the best elements $g_1^*, g_2^*(y_1), \cdots, g_n^*(y_1, \cdots, y_{n-1}), \cdots$, in (2.2). Are the best elements $g_i^*(y_1, \cdots, y_{i-1})$ independent of $y_1, \cdots, y_{i-1}$? (Equivalently, is the optimal information nonadaptive?)

3. Optimal algorithm. In this section we deal with problem (i) of § 2. We describe the μ-spline algorithm $\bar{\varphi}^s = \{\varphi_n^s\}$ using given information $\bar{N} = \{N_n\}$ (see [10], [11]) and show its optimality.

Let $\bar{N}$ be given by (2.2) and (2.3). For given n and $y = N_n(f)$ define

$$\sigma_n = \sigma_n(y) = \sum_{i=1}^{n} (f, g_i(y)) S_\mu g_i(y) = \sum_{i=1}^{n} y_i S_\mu g_i(y) \tag{3.1}$$

where $g_i(y) = g_i(y_1, \cdots, y_{i-1})$. The element σ_n is the unique solution of the problem

$$\begin{aligned} &N_n(\sigma_n) = N_n(f), \\ &\|S_\mu^{-1/2}\sigma_n\| = \inf\{\|S_\mu^{-1/2}g\|\colon g \in S_\mu^{1/2}(F_1),\ N_n(g) = N_n(f)\}. \end{aligned} \tag{3.2}$$

An element σ_n that is the solution of (3.2) is often called a $S_\mu^{-1/2}$-spline. To stress the dependence on the measure μ, we shall call σ_n a *μ-spline*. The algorithm $\bar{\varphi}^s = \{\varphi_n^s\}$, where

$$\varphi_n^s(N_n f) = S\sigma_n(N_n f) = \sum_{i=1}^{n} (f, g_i(y))SS_\mu g_i(y), \tag{3.3}$$

is called a *μ-spline algorithm.*

It is known (see [10], [11]) that the μ-spline algorithm φ_n^s is a unique algorithm that minimizes the average error $\int_{F_1} \|Sf - \varphi_n(N_n(f))\|^2 \mu(df)$ over all measurable mappings $\varphi_n\colon \mathbf{R}^n \to F_2$. We are ready to prove the optimality of $\bar{\varphi}^s$ in the sense of (2.6).

THEOREM 3.1. *The μ-spline algorithm $\bar{\varphi}^s$ is optimal. That is, for any algorithm $\bar{\varphi} = \{\varphi_n\}$ that uses $\bar{N} = \{N_n\}$ we have*

$$\mu\left(\left\{f \in F_1\colon \lim_n \frac{\|Sf - \varphi_n(N_n(f))\|}{\|Sf - \varphi_n^s(N_n(f))\|} = 0\right\}\right) = 0. \tag{3.4}$$

Proof. Given such an algorithm $\bar{\varphi}$, let $A = \{f \in F_1\colon \lim_n \|Sf - \varphi_n(N_n(f))\| / \|Sf - \varphi_n^s(N_n(f))\| = 0\}$. Take a number $q \in (0, 1)$ and define

$$A_n = \{f \in F_1\colon \|Sf - \varphi_n(N_n(f))\| < q\|Sf - \varphi_n^s(N_n(f))\|\}.$$

Then $A \subset \bigcup_{i=1}^{\infty} \bigcap_{n=i}^{\infty} A_n$. Note that A and A_n are measurable and

$$\mu(A) \leqq \lim_i \mu\left(\bigcap_{n=i}^{\infty} A_n\right) \leqq \overline{\lim_n}\, \mu(A_n). \tag{3.5}$$

We estimate $\mu(A_n)$. Define the probability measure

$$\mu_1(B) = \mu(N_n^{-1}B) = \mu(\{f \in F_1\colon N_n(f) \in B\})$$

where B is a Borel set of $\mathbf{R}^n$. (Although it is not needed here, we remark that μ_1 is a Gaussian measure with mean element zero and the identity covariance operator; see [9, Thm. 3.1(i)].)

From Theorem 8.1 of [5, p. 147] we know that there exists a unique (modulo a set of μ_1-measure zero) family of probability measures $\mu_2(\cdot|y)$ defined on Borel sets of F_1 such that

$$\begin{aligned} &\mu_2(N_n^{-1}(y)|y) = 1 \quad \forall y \in \mathbf{R}^n \text{ a.e.}, \\ &\mu_2(B|\cdot) \text{ is } \mu_1\text{-measurable}, \\ &\mu(B) = \int_{\mathbf{R}^n} \mu_2(B|y)\mu_1(df) \end{aligned} \tag{3.6}$$

for any Borel set B of F_1. Thus we have

$$\mu(A_n) = \int_{\mathbf{R}^n} \mu_2(A_n|y)\mu_1(dy). \tag{3.7}$$

It is shown in Theorem 3.1(ii) of [9], that $\mu_2(\cdot|y)$ is a Gaussian measure with mean element $\sigma_n(y)$ and the correlation operator

$$S_y = (I - \sigma_{n,y})S_\mu(I - \sigma_{n,y}^*),$$

where $\sigma_{n,y}: F_1 \to F_1$ is a linear operator such that

$$\sigma_{n,y}(h) = \sum_{i=1}^{n} (h, g_i(y)) S_\mu g_i(y) \quad \forall h \in F_1.$$

For $y = N_n(f)$, let $g_n(y) = \varphi_n(y) - \varphi_n^s(y)$. Then

$$N_n^{-1}(y) \cap A_n = N_n^{-1}(y) \cap B_n(y) + \sigma_n(y)$$

where $B_n(y) = \{h \in F_1: \|Sh - g_n(y)\| < q\|Sh\|\}$. Due to (3.6) and the fact that $\sigma_n(y)$ is the mean element of $\mu_2(\cdot|y)$ we have

$$\mu_2(A_n|y) = \mu_2(N_n^{-1}(y) \cap A_n|y) = \nu_y(B_n(y)) \tag{3.8}$$

where ν_y is a Gaussian measure with mean element zero and covariance operator S_y.

Observe that $S^* g_n(y) = 0$ implies that $B_n(y) = \varnothing$. Assume therefore that $S^* g_n(y) \neq 0$. Let $e_1 = g_n(y)/\|g_n(y)\|$. Every element $z \in F_2$ can be decomposed as $z = (z, e_1)e_1 + z_2$ where $(z_2, e_1) = 0$. Then for $h \in B_n(y)$ we have

$$(Sh - g_n(y), e_1)^2 + \|(Sh)_2\|^2 < q^2 (Sh, e_1)^2 + q^2 \|(Sh)_2\|^2.$$

This yields

$$(1-q^2)(h, S^* e_1)^2 - 2(h, S^* e_1)\|g_n(y)\| + \|g_n(y)\|^2 < 0,$$

and consequently

$$\frac{\|g_n(y)\|}{1+q} < (h, S^* e_1) < \frac{\|g_n(y)\|}{1-q}. \tag{3.9}$$

Since ν_y is a Gaussian measure, (3.9) yields

$$\nu_y(B_n(y)) \leqq \nu_y(\{h: (h, S^* e_1)/\|g_n(y)\| \in ((1+q)^{-1}, (1-q)^{-1})\})$$
$$= \begin{cases} 0 & \text{if } S_y S^* e_1 = 0, \\ \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{a_n}^{b_n} e^{-t^2/2}\, dt & \text{if } S_y S^* e_1 \neq 0, \end{cases}$$

where $a_n = \|g_n(y)\|/((1+q)\sqrt{(S_y S^* e_1, S^* e_1)})$ and $b_n = \|g_n(y)\|/((1-q)\times \sqrt{(S_y S^* e_1, S^* e_1)})$. If $S_y S^* e_1 \neq 0$, we estimate $e^{-t^2/2}$ by its value at a_n. Then we get

$$\nu_y(B_n(y)) \leqq \frac{1}{\sqrt{2\pi}} e^{-a_n^2/2}(b_n - a_n) = \sqrt{\frac{2}{\pi}} \frac{q}{1-q} a_n e^{-a_n^2/2}.$$

Since $x e^{-x^2/2} \leqq 1/\sqrt{e}$ we finally get

$$\nu_y(B_n(y)) \leqq cq/(1-q) \tag{3.10}$$

where $c = \sqrt{2/(\pi e)}$. From (3.7), (3.8) and (3.10) we have $\mu(A_n) \leqq cq/(1-q)$. Then (3.5) yields

$$\mu(A) \leqq cq/(1-q). \tag{3.11}$$

Since q can be arbitrarily small, $\mu(A) = 0$ as claimed. □

Remark 3.1. Theorem 3.1 remains true for more general measures μ. Namely, assume that μ is *elliptically contoured* with mean element zero and covariance operator S_μ (see Crawford [1]). That is, μ is of the form

$$\mu(B) = \int_0^\infty \mu_G\left(\frac{1}{\sqrt{t}} B\right) \alpha(dt) \qquad \forall B\text{-Borel set of } F_1, \tag{3.12}$$

where α is a measure defined on Borel sets of $(0, +\infty)$ such that

$$\int_0^\infty \alpha(dt) = \int_0^\infty t\alpha(dt) = 1. \tag{3.13}$$

Here μ_G denotes the Gaussian measure with mean element zero and covariance operator S_μ. Note that $\nu(B) = \mu_G(1/\sqrt{t}\, B)$ is a Gaussian measure with mean element zero and covariance operator tS_μ. Since (3.11) holds for any Gaussian measure with mean zero, then

$$\mu_G\left(\frac{1}{\sqrt{t}} A\right) \leqq \frac{cq}{1-q}, \qquad c = \sqrt{2/(\pi e)},$$

and consequently (3.12) and (3.13) yield

$$\mu(A) = \int_0^\infty \mu_G\left(\frac{1}{\sqrt{t}} A\right)\alpha(dt) \leqq \frac{cq}{1-q}.$$

This implies that $\mu(A) = 0$ as claimed.

Remark 3.2. The proof of Theorem 3.1 supplies a slightly stronger result than (3.4). Namely, for $q \in (0, 1)$ let

$$B = \left\{ f \in F_1 : \overline{\lim_n} \frac{\|Sf - \varphi_n(N_n(f))\|}{\|Sf - \varphi_n^s(N_n(f))\|} < q \right\}.$$

Then repeating the proof of Theorem 3.1 for the set B instead of the set A, we can obtain

$$\mu(B) \leqq \min\left\{\frac{1}{2}, \sqrt{\frac{2}{\pi e}} \frac{q}{1-q}\right\}. \tag{3.14}$$

Thus for small q, the measure of B is also small. We do not know whether the estimate (3.14) is sharp. Due to Remark 3.1, (3.14) also holds for an elliptically contoured measure.

4. Rate of convergence. In this section we deal with problem (ii) of § 2. We characterize the rate of convergence of the μ-spline algorithm $\bar{\varphi}^s = \{\varphi_n^s\}$ that uses given information $\bar{N} = \{N_n\}$. Observe that

$$\int_{F_1} \|Sf - \varphi_n^s(N_n(f))\|^2 \mu(df) = \int_{\mathbf{R}^n} \int_{F_1} \|Sf - \varphi_n^s(y)\|^2 \mu_2(df|y)\mu_1(dy) \tag{4.1}$$

where μ_1 and $\mu_2(\cdot|y)$ are defined as in § 3. As in [9], we define the local (average) radius rad (N_n, y) of information N_n by

$$\begin{aligned} \text{rad}\,(N_n, y) &= \inf_{g \in F_2} \left\{ \int_{F_1} \|Sf - g\|^2 \mu_2(df|y) \right\}^{1/2} \\ &= \left\{ \int_{F_1} \|Sf - \varphi_n^s(y)\|^2 \mu_2(df|y) \right\}^{1/2} \\ &= \left\{ \int_{F_1} \|Sf\|^2 \nu_y(df) \right\}^{1/2}, \end{aligned} \tag{4.2}$$

where ν_y is a Gaussian measure with mean element zero and covariance operator S_y,

$$S_y h = S_\mu h - \sum_{i=1}^n (h, S_\mu g_i) S_\mu g_i \quad \forall h \in F_1. \tag{4.3}$$

Here $g_i = g_i(y)$, $i = 1, 2, \cdots, n$.

We shall prove that the sequence of local radii rad $(N_n, N_n(f))$ characterizes the rate of convergence of the μ-spline algorithm. Before proving this, we obtain a more explicit form of rad (N_n, y). Let

$$\eta_i = \eta_i(y) = S_\mu^{1/2} g_i(y), \qquad i = 1, 2, \cdots, n. \tag{4.4}$$

Then (2.3) yields $(\eta_i, \eta_j) = \delta_{i,j}$. Define the operator

$$K = S_\mu^{1/2} S^* S S_\mu^{1/2} : F_1 \to F_1. \tag{4.5}$$

Note that $K = K^* \geqq 0$. Furthermore K has finite trace. Indeed, let $\{f_j\}$ be the orthonormal basis of F_1 such that $S_\mu f_j = \lambda_j f_j$. Then

$$\begin{aligned} \text{trace}\,(K) &= \sum_{j=1}^{\infty} (Kf_j, f_j) = \sum_{j=1}^{\infty} \lambda_j \|Sf_j\|^2 \\ &\leqq \|S\|^2 \,\text{trace}\,(S_\mu) < +\infty. \end{aligned} \tag{4.6}$$

Observe also that

$$\text{trace}\,(K) = \int_{F_1} \|Sf\|^2 \mu(df).$$

Indeed, $\|Sf\|^2 = \sum_{i,j=1}^{\infty} (f, f_i)(f, f_j)(Sf_i, Sf_j)$ and

$$\begin{aligned} \int_{F_1} \|Sf\|^2 \mu(df) &= \sum_{i,j=1}^{\infty} (S_\mu f_i, f_j)(Sf_i, Sf_j) \\ &= \sum_{j=1}^{\infty} \lambda_j (S^* S f_j, f_j) = \sum_{j=1}^{\infty} (S^* S S_\mu^{1/2} f_j, S_\mu^{1/2} f_j) \\ &= \sum_{j=1}^{\infty} (Kf_j, f_j) \end{aligned}$$

as claimed.

LEMMA 4.1.

$$\begin{aligned} \text{rad}\,(N_n, y) &= \sqrt{\text{trace}\,(S S_y S^*)} \\ &= \sqrt{\text{trace}\,(K) - \textstyle\sum_{i=1}^{n} (K\eta_i(y), \eta_i(y))}. \end{aligned} \tag{4.7}$$

Proof. Let $\beta_y = \nu_y S^{-1}$ be a measure defined on Borel sets of F_2. Note that β_y is a Gaussian measure defined on Borel sets of F_2 with mean element zero and covariance operator $S_{\beta_y} = S S_y S^*$. Change variables in (4.2) by setting $g = Sf$. Then

$$\begin{aligned} \text{rad}^2\,(N_n, y) &= \int_{F_1} \|Sf\|^2 \nu_y(df) = \int_{F_2} \|g\|^2 \beta_y(dg) \\ &= \text{trace}\,(S_{\beta_y}) = \text{trace}\,(S S_y S^*), \end{aligned}$$

which proves the first equality in (4.7).

To prove the second equality in (4.7), take any orthonormal basis $\{h_j\}$ of F_2. Then (4.3) yields

$$\begin{aligned} \text{trace}\,(S S_y S^*) &= \sum_{j=1}^{\infty} (S S_y S^* h_j, h_j) \\ &= \sum_{j=1}^{\infty} (S S_\mu S^* h_j, h_j) - \sum_{i=1}^{n} \sum_{j=1}^{\infty} (S S_\mu g_i, h_j)^2 \\ &= \text{trace}\,(S S_\mu S^*) - \sum_{i=1}^{n} \|S S_\mu^{1/2} \eta_i\|^2 \\ &= \text{trace}\,(S S_\mu S^*) - \sum_{i=1}^{n} (K\eta_i, \eta_i). \end{aligned}$$

Thus, it is enough to show that trace $(K) = \text{trace}\,(SS_\mu S^*)$. To do this observe that $S^*h_i = \sum_{j=1}^{\infty} (S^*h_i, f_j) f_j$ where $S_\mu f_j = \lambda_j f_j$. Then

$$\begin{aligned}\text{trace}\,(SS_\mu S^*) &= \sum_{i=1}^{\infty} (SS_\mu S^* h_i, h_i) \\ &= \sum_{j=1}^{\infty} \lambda_j \sum_{i=1}^{\infty} (h_i, Sf_j)^2 \\ &= \sum_{j=1}^{\infty} \lambda_j \|Sf_j\|^2 \\ &= \text{trace}\,(K)\end{aligned}$$

due to (4.6). This completes the proof. □

We need an estimate of a Gaussian measure ν of the ball $B_r = \{g \in F_2 : \|g\| \leqq r\}$. We assume that the Gaussian measure ν defined on Borel sets of F_2 has mean element zero and its covariance operator S_ν is nonzero.

THEOREM 4.1 (Kwapień [4]).

$$\nu(B_r) \leqq \tfrac{4}{3}\, \psi(2r/\sqrt{\text{trace}\,(S_\nu)}), \tag{4.8}$$

where

$$\psi(x) = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\tau^2/2}\, d\tau.$$

Proof. Assume first that $S_\nu > 0$. Let $\{\zeta_i\}$ be the orthonormal basis of F_2 such that $S_\nu \zeta_i = \lambda_i \zeta_i$, where $\lambda_i > 0$ and $\text{trace}\,(S_\nu) = \sum_{i=1}^{\infty} \lambda_i$. Define the random variables $\xi_i(g) = (g, \zeta_i)/\sqrt{\lambda_i}$, $i = 1, 2, \cdots$. Then $\{\xi_i\}$ is a sequence of independent random variables each of them with Gaussian distribution with mean zero and variance one. Note that

$$\nu(B_r) = \nu\left(\left\{g \in F_2 : \sum_{i=1}^{\infty} \lambda_i \xi_i^2(g) \leqq r^2\right\}\right).$$

Let λ denote the Lebesgue measure on $[0, 1]$. Let $\{r_i\}$ be the Radamacher system on $[0, 1]$, i.e., $r_i : [0, 1] \to \mathbf{R}$, and $\{r_i\}$ is a sequence of independent random variables each of them with distribution $\lambda(\{t : r_i(t) = +1\}) = \lambda(\{t : r_i(t) = -1\}) = \frac{1}{2}$. For $g \in B_r$ and $c > 0$ we have

$$\lambda\left(\left\{t : \left|\sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \xi_i(g) r_i(t)\right| \leqq c\right\}\right) \geqq R\left(\frac{c}{r}\right), \tag{4.9}$$

where

$$R(x) = \inf_{\sum_{i=1}^{\infty} c_i^2 \leqq 1} \lambda\left(\left\{t : \left|\sum_{i=1}^{\infty} c_i r_i(t)\right| \leqq x\right\}\right). \tag{4.10}$$

Due to Fubini's theorem, we get from (4.9)

$$\begin{aligned}&\nu \otimes \lambda\left(\left\{(g, t) : \left|\sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \xi_i(g) r_i(t)\right| \leqq c\right\}\right) \\ &\quad \geqq \int_{B_r} \lambda\left(\left\{t : \left|\sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \xi_i(g) r_i(t)\right| \leqq c\right\}\right) \nu(dg) \\ &\quad \geqq R\left(\frac{c}{r}\right) \nu(B_r),\end{aligned} \tag{4.11}$$

where $\nu \otimes \lambda$ denotes the usual product measure. On the other hand, let $\zeta_i(g, t) = \xi_i(g) r_i(t)$ for $g \in F_2$ and $t \in [0, 1]$. Then $\{\zeta_i\}$ is a sequence of independent random variables each with Gaussian distribution with mean zero and variance one. Therefore $\Sigma_{i=1}^{\infty} \sqrt{\lambda_i}\, \zeta_i$ has Gaussian distribution with mean zero and variance $\sigma = \sum_{i=1}^{\infty} \lambda_i =$ trace (S_ν). Hence the left-hand side of (4.11) is equal to

$$\frac{1}{\sqrt{2\pi\sigma}} \int_{-c}^{+c} e^{-\tau^2/(2\sigma)}\, d\tau = \sqrt{\frac{2}{\pi}} \int_0^{c/\sqrt{\sigma}} e^{-\tau^2/2}\, d\tau = \psi\left(\frac{c}{\sqrt{\sigma}}\right).$$

Thus we get

$$\nu(B_r) \leqq \psi\left(\frac{c}{\sqrt{\sigma}}\right) \Big/ R\left(\frac{c}{r}\right). \tag{4.12}$$

Let $c = 2r$. To estimate $R(2)$ we use Chebyshev's inequality, which states that

$$\lambda\left(\left\{t: \left|\sum_{i=1}^{\infty} c_i r_i(t)\right| > 2\right\}\right) \leqq \frac{1}{4} \int_0^1 \left(\sum_{i=1}^{\infty} c_i r_i(t)\right)^2 dt.$$

Since r_i are independent with mean zero, then

$$\int_0^1 \left(\sum_{i=1}^{\infty} c_i r_i(t)\right)^2 dt = \sum_{i=1}^{\infty} c_i^2 \leqq 1.$$

Hence $R(2) \geqq 1 - \frac{1}{4} = \frac{3}{4}$ and (4.12) yields (4.8) in the nonsingular case.

Assume now that S_ν is singular and let $X = \ker S_\nu$. Decompose F_2 as the direct sum $X \oplus X^\perp$ where $X^\perp$ is the orthonormal complement of X. Then for every g from F_2, we have $g = g_1 + g_2$, $g_1 \in X$ and $g_2 \in X^\perp$. We get

$$\nu(B_r) = \bar{\nu}(\{h \in X^\perp: \|h\| \leqq r\})$$

where $\bar{\nu}$ is the Gaussian measure on $X^\perp$ with mean element zero and covariance operator $S_\nu|_{X^\perp} > 0$. Applying (4.8) to $\bar{\nu}$, we get the desired estimate on $\nu(B_r)$. Hence the proof is complete. □

We are ready to characterize the rate of convergence of the μ-spline algorithm.

THEOREM 4.2.

$$\mu\left(\left\{f \in F_1: \lim_n \frac{\|Sf - \varphi_n^s(N_n(f))\|}{\operatorname{rad}(N_n, N_n(f))} = 0\right\}\right) = 0. \tag{4.13}$$

Proof. Let $A = \{f \in F_1: \lim_n \|Sf - \varphi_n^s(N_n(f))\| / \operatorname{rad}(N_n, N_n(f)) = 0\}$. Take a number $q \in (0, 1)$ and define

$$A_n = \{f \in F_1: \|Sf - \varphi_n^s(N_n(f))\| < q \operatorname{rad}(N_n, N_n(f))\}.$$

Then $A \subset \bigcup_{i=1}^{\infty} \bigcap_{n=i}^{\infty} A_n$ and $\mu(A) \leqq \overline{\lim_n}\, \mu(A_n)$. We estimate $\mu(A_n)$. From (3.7) we have

$$\begin{aligned} \mu(A_n) &= \int_{\mathbf{R}^n} \mu_2(\{f: \|Sf - \varphi_n^s(y)\| < q \operatorname{rad}(N_n, y)\} | y) \mu_1(dy) \\ &= \int_{\mathbf{R}^n} \nu_y(\{f: \|Sf\| < q \operatorname{rad}(N_n, y)\}) \mu_1(dy), \end{aligned} \tag{4.14}$$

where ν_y is the Gaussian measure with mean zero and covariance operator S_y given by (4.3). As in the proof of Lemma 4.1, let $\beta_y = \nu_y S^{-1}$ be the Gaussian measure with

mean element zero and covariance operator $S_{\beta_y} = SS_yS^*$. Let $r = q \operatorname{rad}(N_n, y)$ and $B_r = \{g \in F_2 : \|g\| \leq r\}$. Then

$$\beta_y(B_r) = \nu_y(\{f \in F_1 : \|Sf\| < r\}).$$

Due to Lemma 4.1 we have $r = q\sqrt{\operatorname{trace}(S_{\beta_y})}$. From (4.14) we get

$$\mu(A_n) = \int_{\mathbf{R}^n} \beta_y(\{g \in F_2 : \|g\| < q\sqrt{\operatorname{trace}(S_{\beta_y})}\})\mu_1(dy). \tag{4.15}$$

From Theorem 4.1 we have the following estimate:

$$\mu(A_n) \leq \frac{4}{3}\int_{\mathbf{R}^n} \psi(2q)\mu_1(dy) = \frac{4}{3}\psi(2g),$$

and consequently

$$\mu(A) \leq \tfrac{4}{3}\psi(2q).$$

Since q can be arbitrarily small and $\psi(2q)$ tends to zero with q, we conclude $\mu(A) = 0$, as claimed. □

Remark 4.1. It is also true that (4.13) of Theorem 4.2 holds for any algorithm $\bar{\varphi} = \{\varphi_n\}$ using $\bar{N} = \{N_n\}$. That is,

$$\mu\left(\left\{f \in F_1 : \lim_n \frac{\|Sf - \varphi_n(N_n(f))\|}{\operatorname{rad}(N_n, N_n(f))} = 0\right\}\right) = 0.$$

Indeed, repeating the proof of Theorem 4.2 we get for $B = \{f \in F_1 : \lim_n \|Sf - \varphi_n(N_n(f))\|/\operatorname{rad}(N_n, N_n(f)) = 0\}$

$$\mu(B) \leq \overline{\lim_n} \int_{\mathbf{R}^n} \beta_y(\{g \in F_2 : \|g - a\| < q\sqrt{\operatorname{trace}(S_{\beta_y})}\})\mu_1(dy),$$

where $a = a(y) = \varphi_n(y) - \varphi_n^s(y)$. It is known (see for instance [9]) that a Gaussian measure of the ball $B_r(a)$ of radius r and center a is maximal for $a = 0$. Thus

$$\begin{aligned}\beta_y(\{g \in F_2 : \|g - a\| < q\sqrt{\operatorname{trace}(S_{\beta_y})}\}) &\leq \beta_y(\{g \in F_2 : \|g\| < q\sqrt{\operatorname{trace}(S_{\beta_y})}\}) \\ &\leq \tfrac{4}{3}\psi(2q),\end{aligned}$$

due to Theorem 4.1. Therefore $\mu(B) \leq \frac{4}{3}\psi(2q)$, and since q can be arbitrarily small, $\mu(B) = 0$, as claimed.

Remark 4.2. Theorem 4.2 remains true for more general measures μ. Namely, assume as in Remark 3.1 that μ is elliptically contoured. Then for

$$A = \left\{f \in F_1 : \lim_n \frac{\|Sf - \varphi_n^s(N_n(f))\|}{\sqrt{\operatorname{trace}(SS_{N_n(f)}S^*)}} = 0\right\},$$

we have

$$\mu(A) = 0.$$

To prove this let A_n be defined as in the proof of Theorem 4.2. Then

$$\mu(A) \leq \mu(A_n) = \int_0^\infty \mu_G\left(\frac{1}{\sqrt{t}} A_n\right)\alpha(dt).$$

For every fixed t, $\mu_G(1/\sqrt{t}\,\cdot)$ is a Gaussian measure with mean element zero and covariance operator tS_μ. Repeating the proof of Theorem 4.2, we therefore obtain that

$$\mu(A_n) \leqq \frac{4}{3}\int_0^\infty \int_{\mathbf{R}^n} \psi(2q/\sqrt{t})\mu_1(dy)\alpha(dt) = \frac{4}{3}\int_0^\infty \psi(2g/\sqrt{t})\alpha(dt).$$

From the definition of ψ we have

$$\begin{aligned}\int_0^\infty \psi(2q/\sqrt{t})\alpha(dt) &= \sqrt{\frac{2}{\pi}}\int_0^\infty \int_0^{2q/\sqrt{t}} e^{-\tau^2/2}d\tau\,\alpha(dt)\\ &= \sqrt{\frac{2}{\pi}}\int_0^q \int_0^{2q/\sqrt{t}} e^{-\tau^2/2}\,d\tau\alpha(dt) + \sqrt{\frac{2}{\pi}}\int_q^\infty \int_0^{2q/\sqrt{t}} e^{-\tau^2/2}\,d\tau\,\alpha(dt)\\ &\leqq \int_0^q \left\{\sqrt{\frac{2}{\pi}}\int_0^\infty e^{-\tau^2/2}\,d\tau\right\}\alpha(dt) + \sqrt{\frac{2}{\pi}}\int_q^\infty \int_0^{2\sqrt{q}} e^{-\tau^2/2}\,d\tau\,\alpha(dt)\\ &= \alpha((0,q]) + \sqrt{\frac{2}{\pi}}\int_0^{2\sqrt{q}} e^{-\tau^2/2}d\tau \overset{df}{=} a(q).\end{aligned}$$

Since $0 = \alpha(\varnothing) = \lim_{q\to 0^+}\alpha((0,q])$, we see that $\alpha((0,q])$ and $\sqrt{2/\pi}\int_0^{2\sqrt{q}} e^{-\tau^2/2}d\tau$ tend to zero with q. Hence $\lim_{q\to 0} a(q) = 0$. Since $\mu(A) \leqq \mu(A_n) \leqq a(q)$ and q can be arbitrarily small, $\mu(A) = 0$, as claimed.

Theorem 4.2 states that modulo a set of measure zero the μ-spline algorithm does not converge *faster* than the sequence of local radii. We now show that the μ-spline algorithm does not converge more *slowly* than the sequence of local radii.

THEOREM 4.3.

$$\mu\left(\left\{f \in F_1 : \lim_n \frac{\operatorname{rad}(N_n, N_n(f))}{\|Sf - \varphi_n^s(N_n(f))\|} = 0\right\}\right) = 0. \tag{4.16}$$

Proof. Let $A = \{f \in F_1 : \lim_n \operatorname{rad}(N_n, N_n(f))/\|Sf - \varphi_n^s(N_n f)\| = 0\}$. Take a number $q \in (0,1)$ and define

$$A_n = \left\{f \in F_1 : \|Sf - \varphi_n^s(N_n(f))\| \geqq \frac{1}{q}\operatorname{rad}(N_n, N_n(f))\right\}.$$

Then $A \subset \bigcup_{i=1}^\infty \bigcap_{n=i}^\infty A_n$ and $\mu(A) \leqq \overline{\lim}_n \mu(A_n)$. As in the proof of Theorem 4.2, we conclude that

$$\mu(A_n) = 1 - \int_{\mathbf{R}^n} \beta_y\left(\left\{g \in F_2 : \|g\| < \frac{1}{q}\sqrt{\operatorname{trace}(S_{\beta_y})}\right\}\right)\mu_1(dy). \tag{4.17}$$

For any probability measure ν with covariance operator S_ν and for any ball $B_r = \{g \in F_2 : \|g\| < r\}$, we have

$$\begin{aligned}\operatorname{trace}(S_\nu) &= \int_{F_2} \|g\|^2 \nu(dg) \geqq \int_{F_2 - B_r} \|g\|^2 \nu(dg) \geqq r^2 \nu(F_2 - B_r)\\ &= r^2(1 - \nu(B_r)).\end{aligned}$$

Thus

$$\nu(B_r) \geqq 1 - \frac{\operatorname{trace}(S_\nu)}{r^2}.$$

In particular,

$$\beta_y\left(\left\{g \in F_2 : \|g\| < \frac{1}{q}\sqrt{\operatorname{trace}(S_{\beta_y})}\right\}\right) \geqq 1 - q^2.$$

From (4.17) we get $\mu(A_n) \leqq q^2$, and consequently $\mu(A) \leqq q^2$. Since q can be arbitrarily small, $\mu(A) = 0$, as claimed. □

Remark 4.3. The proof of Theorem 4.3 supplies a slightly stronger result than (4.12). Namely for $q \in (0, 1)$ let

$$B = \left\{ f \in F_1 : \overline{\lim_n} \frac{\operatorname{rad}(N_n, N_n(f))}{\|Sf - \varphi_n^s(N_n(f))\|} \leqq q \right\}.$$

Then repeating the proof of Theorem 4.3, we get

$$\mu(B) \leqq q^2. \tag{4.18}$$

Thus for small q, $\mu(B)$ is close to zero. We do not know whether the estimate (4.18) is sharp.

5. Optimal information. In this section we deal with problem (iii) of § 2. We find optimal information $\bar{N}^* = \{N_n^*\}$ of the form (2.2) for which the rate of convergence of the μ-spline algorithm $\bar{\varphi}^s = \{\varphi_n^s\}$ using $\bar{N}^*$ is best possible. Due to the result of § 4 this is equivalent to finding information $\bar{N}^*$ for which the sequence of local radii $\operatorname{rad}(N_n^*, N_n^*(f))$ goes to zero as fast as possible.

As in § 4, let

$$K = S_\mu^{1/2} S^* S S_\mu^{1/2} : F_1 \to F_1. \tag{5.1}$$

We know that $K = K^* \geqq 0$ and K has finite trace. Let m denote the total number of positive eigenvalues of K. Observe that m can be infinite. Let $\{\eta_i^*\}$, $i < m+1$, be the orthonormal eigenelements of K,

$$K\eta_i^* = \lambda_i^* \eta_i^*, \qquad \lambda_1^* \geqq \lambda_2^* \geqq \cdots > 0. \tag{5.2}$$

For $i < m+1$ define $g_i^* = S_\mu^{-1/2} \eta_i^*$. Then $(S_\mu g_i^*, g_j^*) = (\eta_i^*, \eta_j^*) = \delta_{i,j}$. For $n < m+1$ define

$$N_n^*(f) = [(f, g_1^*), (f, g_2^*), \cdots, (f, g_n^*)]. \tag{5.3}$$

Note that N_n^* is nonadaptive. Its local radius $\operatorname{rad}(N_n^*, y)$ given by (4.7) and (4.4) is independent of y and equal to

$$\operatorname{rad}(N_n^*) = \operatorname{rad}(N_n^*, y) = \sqrt{\sum_{i=n+1}^{m} \lambda_i^*}. \tag{5.4}$$

If m is finite then $\operatorname{rad}(N_m^*) = 0$ and $Sf = \varphi_m^s(N_m^*(f))$, for all $f \in F_1$. This means that we approximate Sf exactly for any f using m inner products. Therefore without loss of generality we assume from now on that $m = +\infty$. Define the information

$$\bar{N}^* = \{N_n^*\} \tag{5.5}$$

where N_n^* is given by (5.3). We stress that $\bar{N}^*$ is *nonadaptive* and the μ-spline algorithm $\bar{\varphi}^s = \{\varphi_n^s\}$ that uses $\bar{N}^*$ has the form

$$\varphi_n^s(N_n^*(f)) = \sum_{i=1}^{n} (f, \eta_i^*) S\eta_i^*.$$

Hence $\bar{\varphi}^s$ is *linear.*

Theorem 4.3 states that the μ-spline algorithm $\bar{\varphi}^s$ converges at least as fast as the sequence of the radii $\{\operatorname{rad}(N_n^*)\}$, i.e.,

$$\mu\left(\left\{ f \in F_1 : \lim_n \frac{\operatorname{rad}(N_n^*)}{\|Sf - \varphi_n^s(N_n^*(f))\|} = 0 \right\} \right) = 0.$$

We are ready to prove the optimality of $\bar{N}^*$.

THEOREM 5.1. *The nonadaptive information operator $\bar{N}^*$ is optimal in the class of adaptive information operators of the form* (2.2), *i.e., for any adaptive information operator $\bar{N}=\{N_n\}$ and any algorithm $\bar{\varphi}=\{\varphi_n\}$ using $\bar{N}$ we have*

$$\mu\left(\left\{f\in F_1\colon \lim_n \frac{\|Sf-\varphi_n(N_n(f))\|}{\operatorname{rad}(N_n^*)}=0\right\}\right)=0. \tag{5.6}$$

Proof. Let $A=\{f\in F_1\colon \lim_n \|Sf-\varphi_n(N_n(f))\|/\operatorname{rad}(N_n^*)=0\}$ and $B=\{f\in F_1\colon \lim_n \|Sf-\varphi_n(N_n(f))\|/\operatorname{rad}(N_n, N_n(f))=0\}$. It is known (see [10], [11]) that

$$\operatorname{rad}(N_n^*)\leqq \operatorname{rad}(N_n, N_n(f)) \quad \forall f\in F_1.$$

Therefore $A\subset B$ and $\mu(A)\leqq\mu(B)$. Due to Remark 4.1, $\mu(B)=0$. Thus $\mu(A)=0$ as claimed. □

Theorem 5.1 states that adaption does not help for approximation of linear operators in the asymptotic setting of this paper. This agrees with the similar result in a worst case, an average case (see [7], [2], [11]), respectively, and with the results of [8].

The best possible rate of convergence is obtained by the μ-spline algorithm using the information $\bar{N}^*$. This rate of convergence is given by $\operatorname{rad}(N_n^*)=\sqrt{\sum_{i=n+1}^{\infty}\lambda_i^*}$. Thus it depends on how fast the truncated series of the trace of K goes to zero. For instance, if $\lambda_i^*=i^{-r}$ for some $r>1$, then

$$\operatorname{rad}(N_n^*)\cong\frac{1}{r-1}\,\frac{1}{(n+1)^{(r-1)/2}}.$$

If $\lambda_i^*=q^i$ for $q\in(0,1)$, then

$$\operatorname{rad}(N_n^*)=\sqrt{\frac{q^{n+1}}{1-q}}.$$

Appendix. Here we prove the result mentioned in the Introduction. In fact, we will show a slightly stronger result. Namely, we relax the assumptions on the spaces F_1 and F_2 and the solution operator S. In this Appendix we assume that F_1 is an infinite-dimensional Banach space, F_2 is a normed linear space and that S is a one-to-one operator (not necessarily linear).

Let $\bar{\varphi}=\{\varphi_n\}$ be an algorithm using information $\bar{N}=\{N_n\}$. Here we assume that

$$N_n(f)=[L_1(f), L_2(f; y_1), \cdots, L_n(f; y_1, \cdots, y_{n-1})]$$

where $y_1=L_1(f)$, $y_i=L_i(f; y_1, y_2, \cdots, y_{i-1})$ and $L_i(\cdot; y_1, \cdots, y_{i-1})$ is a continuous linear functional. We also assume that $L_1\not\equiv 0$. Define

$$A(\bar{\varphi})=\{f\in F_1\colon \varphi_n(N_n(f))=Sf, \forall n\geqq k(f)\}$$

as the set of elements for which the algorithm $\bar{\varphi}$ solves the problem exactly for sufficiently large n.

LEMMA A.1. *The set $A(\bar{\varphi})$ has empty interior for every algorithm $\bar{\varphi}$.*

Proof. Assume on the contrary that $A(\bar{\varphi})$ contains a ball $B(f_1, r)=\{f\in F_1\colon \|f-f_1\|\leqq r\}$ for $r>0$. We construct a sequence $\{f_j\}$ from the ball $B(f_1, r)$. Suppose inductively that f_j is defined. Since $f_j\in B(f_1, r)\supset A(\bar{\varphi})$, then there exists $k_j=k(f_j)$ such that $\varphi_n(N_n(f_j))=Sf_j$ for $n\geqq k_j$. Take $h_j\in F_1$ such that $L_1(h_j)=L_2(h_j, y_1)=\cdots=L_{k_j}(h_j; y_1, \cdots, y_{k_j-1})=0$ and $\|h_j\|=r/3^j$. Here $y_i=L_i(f_j; y_1, \cdots, y_{i-1})$. Observe that such an element h_j exists since $\dim F_1=+\infty$. Define $f_{j+1}=f_j+h_j$. Then $f_{j+1}=f_1+\sum_{i=1}^{j}h_i$ and $\|f_{j+1}-f_1\|\leqq r\sum_{i=1}^{j}3^{-i}<r/2$. Thus $f_{j+1}\in B(f_1, r)$ and $N_{k_j}(f_{j+1})=N_{k_j}(f_j)$. Hence $\varphi_{k_j}(N_{k_j}(f_{j+1}))=\varphi_{k_j}(N_{k_j}(f_j))=Sf_j\neq Sf_{j+1}$ since $f_{j+1}\neq f_j$ and S is one to one. Thus $k_{j+1}=k(f_{j+1})\geqq k_j+1$.

Define $f = \lim_j f_j = f_1 + \sum_{j=1}^{\infty} h_j$. The element f exists since $\{f_j\}$ is a Cauchy sequence and F_1 is a Banach space. Note that $f \in B(f_1, r)$ and

$$\|f - f_j\| = \left\| \sum_{i=j}^{\infty} h_i \right\| \geq \|h_j\| - \sum_{i=j+1}^{\infty} \|h_i\| = \frac{1}{2} r/3^j > 0.$$

Thus $f \neq f_j$, for all j. The continuity of functionals that form $\bar{N}$ yields

$$N_{k_j}(f) = N_{k_j}\left(f_1 + \sum_{i=1}^{j-1} h_i\right) = N_{k_j}(f_j) \quad \forall j.$$

Thus we have

$$\varphi_{k_j}(N_{k_j}(f)) = \varphi_{k_j}(N_{k_j}(f_j)) = Sf_j \neq Sf.$$

Since k_j goes to infinity with j, this proves that $f \notin A(\bar{\varphi})$. This is a contradiction which completes the proof. □

Remark A.1. In fact, the proof of Lemma A.1 supplies a slightly stronger result. Namely, the proof yields that for every $f \in A(\bar{\varphi})$ and for every $r > 0$, there exists an element h such that $\|h\| \leq r$, $f + h \notin A(\bar{\varphi})$ and additionally $L_1(h) = 0$.

Based on Lemma A.1 and Remark A.1 we prove the result mentioned in the Introduction. Let $\bar{\varphi} = \{\varphi_n\}$ and $\bar{\varphi}^* = \{\varphi_n^*\}$ be two algorithms using information $\bar{N} = \{N_n\}$. Let $a_n(f) = \|Sf - \varphi_n(N_n(f))\|$ and $b_n(f) = \|Sf - \varphi_n^*(N_n(f))\|$. Then

$$A(\bar{\varphi}, \bar{\varphi}^*) = \{f \in F_1 : \lim_n a_n(f)/b_n(f) = 0\}.$$

We need to consider the case $a_n(f) = 0$, for all $n \geq n_0$, for some n_0. Then, if $b_n(f) = 0$, for all $n \geq n_1$, for some n_1, the two algorithms $\bar{\varphi}$ and $\bar{\varphi}^*$ solve the problem exactly and none of them are superior. It is therefore reasonable to set in this case $\lim_n a_n(f)/b_n(f) = 1$. On the other hand, if $\{b_n(f)\}$ contains a nonzero subsequence, then the algorithm $\bar{\varphi}$ is superior to the algorithm $\bar{\varphi}^*$ and it is reasonable to set in this case $\lim_n a_n(f)/b_n(f) = 0$. Having this convention in mind we are ready to prove the following theorem.

THEOREM A.1. *For every algorithm $\bar{\varphi}^*$ there exists an algorithm $\bar{\varphi}$ such that the set $A(\bar{\varphi}, \bar{\varphi}^*)$ is uncountable.*

Proof. Choose an element $f_1 \in F_1$ such that $L_1(f_1) = 1$. For $y \in \mathbf{R}$, define

$$g_y = \begin{cases} yf_1 & \text{if } yf_1 \notin A(\bar{\varphi}^*), \\ yf_1 + h & \text{if } yf_1 \in A(\bar{\varphi}^*). \end{cases}$$

Here h is chosen such that $L_1(h) = 0$ and $yf_1 + h \notin A(\bar{\varphi}^*)$. Such an element h exists due to Remark A.1.

We now define the algorithm $\bar{\varphi} = \{\varphi_n\}$ as

$$\varphi_n(N_n(f)) = Sg_{N_1(f)}.$$

Note that $\varphi_n(N_n(g_y)) = Sg_y$, for all $n \geq 1$ and $y \in \mathbf{R}$. Since $g_y \notin A(\bar{\varphi}^*)$, there exists a subsequence $\{n_i\}$ such that

$$\varphi_{n_i}^*(N_{n_i}(g_y)) \neq Sg_y.$$

This means that $g_y \in A(\bar{\varphi}, \bar{\varphi}^*)$, for all $g \in \mathbf{R}$. Note that g_y varies with y which means that $A(\bar{\varphi}, \bar{\varphi}^*)$ contains at least as many elements as $\mathbf{R}$. This completes the proof. □

Acknowledgments. We are pleased to thank S. Kwapień for providing the proof of Theorem 4.1 and useful remarks concerning measure theory.

We also thank J. F. Traub, K. Sikorski, and A. G. Werschulz for stimulating discussions and help during the preparation of this paper.

REFERENCES

[1] J. J. CRAWFORD, *Elliptically contoured measures on infinite-dimensional Banach spaces*, Studia Math., 60 (1977), pp. 15-32.

[2] J. B. KADANE, G. W. WASILKOWSKI, AND H. WOŹNIAKOWSKI, *Can adaption help on the average for stochastic information*? J. Complexity, to appear.

[3] HUI-HSUING KUO, *Gaussian Measures in Banach Spaces*, Lecture Notes in Mathematics 463, Springer-Verlag, Berlin, 1975.

[4] S. KWAPIEŃ, private communication, 1984.

[5] K. R. PARATHASARATHY, *Probability Measures on Metric Spaces*, Academic Press, New York, 1967.

[6] A. V. SKOROKHOD, *Integration in Hilbert Space*, Springer-Verlag, New York, 1974.

[7] J. F. TRAUB AND H. WOŹNIAKOWSKI, *A General Theory of Optimal Algorithms*, Academic Press, New York, 1980.

[8] J. M. TROJAN, *Asymptotic setting for linear problems*, unpublished manuscript.

[9] G. W. WASILKOWSKI, *Optimal algorithms for linear problems with Gaussian measure*, Rocky Mountain J. Math., 16 (1986), pp. 727-749.

[10] G. W. WASILKOWSKI AND H. WOŹNIAKOWSKI, *Average case optimality for linear problems in Hilbert spaces*, J. Approx. Theory, 47 (1986), pp. 17-25.

[11] ———, *Can adaption help on the average*?, Numer. Math., 44 (1984), pp. 169-190.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

011

ENTRAINMENT OF A LIMIT-CYCLE OSCILLATOR WITH SHEAR BY LARGE AMPLITUDE FORCING*

MICHAEL ST. VINCENT†

Abstract. The entrainment of a circularly symmetric limit-cycle oscillator due to large amplitude periodic forcing is investigated. The unforced oscillator contains parameters controlling the local strength of attraction of the limit cycle and the amount of radial variation of angular velocity, or shear, near the limit cycle. It is shown that entrainment (1:1 phase locking) will occur for sufficiently large amplitude forcing whenever the local strength of attraction to the limit cycle is great enough. Furthermore, if the amount of shear is allowed to increase with the strength of attraction to the limit cycle, then increasing the ratio of shear to strength of attraction can have the effect of increasing the sensitivity of the oscillator to forcing.

Key words. entrainment, limit cycle, oscillator, phase locking

AMS(MOS) subject classifications. 34C15, 34C25

1. Introduction. This paper is concerned with the periodically forced nonlinear oscillator

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix}=\begin{pmatrix}\lambda & -\omega\\ \omega & \lambda\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}+b_0\begin{pmatrix}u\\ v\end{pmatrix}p(\sigma t) \tag{1.1}$$

defined in a neighborhood of the unit circle, on which it is assumed to have a stable limit cycle. Here "$\cdot$" $= d/dt$, $\lambda=\lambda(r,\varepsilon)$, $\omega=\omega(r,\beta)$, $(r=\sqrt{x^2+y^2})$, $\lambda(1,\varepsilon)=0$, $\omega(1,\beta)=1$, and $p(s)$ is an odd 2π-periodic continuous function satisfying $p(s+\pi)=-p(s)$. All quantities are real, with ε, β, σ and b_0 positive, and $u^2+v^2=1$ (u, v constants).

The unforced oscillator ($b_0=0$) is a circularly symmetric limit-cycle oscillator of "$\lambda-\omega$" form. Oscillators of this form (though not always with a limit cycle on the unit circle) have been used to model a variety of nonlinear phenomena, including those of chemical kinetics and biology, and periodically forced Hopf bifurcations (e.g., [1]-[3]). When written in polar coordinates ($x=r\cos\theta$, $y=r\sin\theta$), it is given by

$$\dot{r}=r\lambda(r,\varepsilon),\qquad \dot{\theta}=\omega(r,\beta).$$

From this we see that λ determines the rate of approach of nearby solutions to the limit cycle, ω determines their angular velocity, and the limit-cycle solutions have constant angular speed 1.

The parameters ε and β are intended to control properties of the oscillator near the limit cycle. More specifically, the dependence on these parameters is meant to be such that the local strength of attraction of the limit cycle becomes infinite as $\varepsilon\to 0$, while the local rate of radial variation in angular velocity, or *shear*, becomes infinite as $\beta\to 0$. To this end, it is assumed that $\lambda_r(1,\varepsilon)\to-\infty$ as $\varepsilon\to 0$ and $|\omega_r(1,\beta)|\to\infty$ as $\beta\to 0$. (Additional assumptions are contained in § 2.)

The main results obtained in this paper concern large amplitude forcing, in contrast with many papers on nonlinear oscillations that consider only small amplitude forcing. In particular, it will be shown that if the forcing amplitude b_0 is sufficiently large, then (1.1) will be entrained, or 1:1 phase locked, whenever the strength of attraction to the

* Received by the editors January 6, 1986; accepted for publication (in revised form) May 21, 1987. This work was supported in part by National Science Foundation grant MCS8301249.

† Department of Mathematics and Computer Science, Clark University, Worcester, Massachusetts 01610.

limit cycle is great enough (i.e., ε is small enough), depending on b_0. In saying that (1.1) is entrained, it is meant that it has a stable periodic solution, of the same period as the forcing, that undergoes one oscillation in each period. It will also be shown that 1:1 phase locking does not occur for small amplitude forcing, except in the resonant case $\sigma = 1$.

A related result for large amplitude forcing of the van der Pol oscillator was obtained by Lloyd [4], and explained heuristically by Levi [5, p. 25]. In that case, both the shear and the strength of attraction to the limit cycle became infinite as a single parameter became small. This will also be considered in (1.1) by allowing β to depend on ε. However, an additional conclusion will be reached here, due to the fact that we can specify the relative magnitudes of the shear and strength of attraction to the limit cycle. In particular, it will be shown that the presence of large shear can reduce the amplitude of forcing required for entrainment. This will be done by showing that if β is made to depend on ε in such a way that $\omega_r(1, \beta)/\lambda_r(1, \varepsilon) \to c$ as $\varepsilon \to 0$, then the size of $b_0\sqrt{1+c^2}$, not just b_0, determines if entrainment will occur for small ε. Consequently, entrainment could be made to occur for any given forcing amplitude by making $|c|$ large enough. In addition, it will be seen how a nonzero value of c affects the location of the periodic solutions in the entrained state.

It is interesting to note that shear, which is shown by this paper to affect entrainment, is also important for the existence of "shock structure" solutions of reaction-diffusion equations [13].

The remainder of this paper is organized as follows: In § 2, (1.1) is simplified by a rotation of coordinates and a change of parameters. It is then shown that there is a family of nested attracting annuli A_ε, each of which is of width $O(\varepsilon)$ and contains the unit circle in its interior. All subsequent analysis of (1.1) concerns the behavior of solutions in A_ε for ε small. In § 3, a further change of variable is made in order to deal with the presence of large shear, and estimates are given for approximating the solutions in A_ε. In particular, it will be shown that the angular component of solutions in A_ε can be approximated by solutions of an equation of the form $\dot{\psi} = 1 - bp(\sigma t)\sin\psi$, which is analyzed in § 4. This phase approximation will be seen to be valid for large amplitude forcing, in addition to the more usual small amplitude case. The results of the previous sections are then used in § 5 to derive the results concerning entrainment. It is shown there that the period map will have exactly two fixed points in A_ε, one a sink and the other a saddle, when entrainment occurs. The paper then concludes with a brief discussion of the effects of large shear.

2. Existence of attracting annuli A_ε. We begin by stating some additional assumptions and simplifying (1.1). Let $I \subset (0, \infty)$ be an open interval containing 1, and let ε_M, β_M and k be positive constants. Then we assume that $\lambda(r, \varepsilon)$ is defined and continuous on $I \times (0, \varepsilon_M)$, $\omega(r, \beta)$ is defined and continuous on $I \times (0, \beta_M)$ and that they are twice continuously differentiable with respect to r. Furthermore, we assume that λ, ω and p satisfy the following:

($\lambda 1$) $\lambda_r(1, \varepsilon)$ is strictly monotonic, and $\lambda_r(1, \varepsilon) \to -\infty$ as $\varepsilon \to 0$;

($\lambda 2$) $|\lambda_{rr}| \leqq k|\lambda_r(1, \varepsilon)|$;

($\omega 1$) $\omega_r(1, \beta)$ is strictly monotonic, and $|\omega_r(1, \beta)| \to \infty$ as $\beta \to 0$;

($\omega 2$) $|\omega_{rr}| \leqq k|\omega_r(1, \beta)|$;

($p1$) $p(s) > 0$ for $s \in (0, \pi)$;

($p2$) $\max |p(s)| = 1$.

An example of functions satisfying these assumptions is provided by $\lambda = (1-r)/\varepsilon$, $\omega = 1+(r-1)/\beta$ and $p = \sin(s)$.

As is clear from the above, (1.1) is defined for all (x, y, t) for which (x, y) is in the annulus $A = \{(x, y) \mid x^2+y^2 = r^2, r \in I\}$. In addition, some simplifications can be made: since λ and ω are unchanged by a rotation of coordinates, rotation through an angle ϕ for which $\cos\phi = u$ and $\sin\phi = v$ allows us to assume that $u = 1$ and $v = 0$. Furthermore, if we let $\bar{\varepsilon} = -1/\lambda_r(1, \varepsilon)$ and $\bar{\beta} = 1/\omega_r(1, \beta)$, then assumptions $(\lambda 1)$ and $(\omega 1)$ allow us to invert these to get continuous functions $\varepsilon(\bar{\varepsilon})$ and $\beta(\bar{\beta})$. As a result, reparametrizing λ and ω by $\bar{\varepsilon}$ and $\bar{\beta}$ shows that we may also assume that $\lambda_r(1, \varepsilon) = -1/\varepsilon$ and $\omega_r(1, \beta) = 1/\beta$. Now, however, it is only assumed that β is of constant sign, with $|\beta| < \beta_M$, since no assumption was made concerning the sign of $\omega_r(1, \beta)$. Thus, from now on we only consider (1.1) in the case

$$\dot{x} = \lambda x - \omega y + b_0 p(\sigma t), \qquad \dot{y} = \omega x + \lambda y, \tag{2.1}$$

with conditions $(\lambda 1-2)$ and $(\omega 1-2)$ replaced by

$$\lambda_r(1, \varepsilon) = -1/\varepsilon, \tag{$\lambda 1'$}$$

$$|\lambda_{rr}| \leqq k/\varepsilon, \tag{$\lambda 2'$}$$

$$\omega_r(1, \beta) = 1/\beta, \tag{$\omega 1'$}$$

$$|\omega_{rr}| \leqq k/|\beta|. \tag{$\omega 2'$}$$

These conditions will now be used to conclude that there is a family of attracting annuli $A_\varepsilon \subset A$ of width $O(\varepsilon)$, each of which contains the unit circle. To begin, use $(\lambda 1')$ and $(\omega 1')$ to write

$$\lambda(r, \varepsilon) = -\frac{1}{\varepsilon}(r-1) + \lambda_1(r, \varepsilon), \qquad \omega(r, \beta) = 1 + \frac{1}{\beta}(r-1) + \omega_1(r, \beta)$$

where

$$\lambda_1 = \int_1^r (r-s)\lambda_{rr}(s, \varepsilon)\, ds \quad \text{and} \quad \omega_1 = \int_1^r (r-s)\omega_{rr}(s, \beta)\, ds.$$

Together with conditions $(\lambda 2')$ and $(\omega 2')$, this immediately yields the following lemma.

LEMMA 2.1. *The following hold on the domains of λ and ω:*

(i) $|\lambda_1(r, \varepsilon)| \leqq k|r-1|^2/\varepsilon$;

(ii) $|\lambda_{1r}(r, \varepsilon)| \leqq k|r-1|/\varepsilon$;

(iii) $|\omega_1(r, \beta)| \leqq k|r-1|^2/|\beta|$;

(iv) $|\omega_{1r}(r, \beta)| \leqq k|r-1|/|\beta|$. □

Now let $I_\varepsilon = [1-2\varepsilon b_0, 1+2\varepsilon b_0]$ and define the annulus A_ε by

$$A_\varepsilon = \{(x, y) \mid x^2+y^2 = r^2, r \in I_\varepsilon\}.$$

For any given b_0, we clearly have $I_\varepsilon \subset I$, and so also $A_\varepsilon \subset A$, whenever ε is small enough. To see that A_ε is attracting for small ε, write (2.1) in polar coordinates $(x = r\cos\eta, y = r\sin\eta)$ to get

$$\dot{r} = r\lambda + b_0 p(\sigma t)\cos\eta, \qquad \dot{\eta} = \omega - (b_0/r)p(\sigma t)\sin\eta. \tag{2.2}$$

Since $|p(\sigma t)| \leqq 1$ from assumption ($p2$), this shows that the vector field will be pointing in on the boundary of A_ε if $r\lambda < -b_0$ for $r = 1+2\varepsilon b_0$ and $r\lambda > b_0$ for $r = 1-2\varepsilon b_0$. We can now prove the following.

THEOREM 2.2. *The annulus $A_\varepsilon \subset A$ will be attracting if $\varepsilon < (1+k-\sqrt{k^2+1})/(4b_0k)$.*

Proof. Let ε be as above, and recall that $\lambda = (-1/\varepsilon)(r-1)+\lambda_1$. From (i) of Lemma 2.1, it follows that $|\lambda_1|/b_0 \leqq 4kb_0\varepsilon < 1+k-\sqrt{k^2+1} < 1$ for $r = 1 \pm 2\varepsilon b_0$. Consequently, on

(1) $r = 1+2\varepsilon b_0$, we have $r\lambda = -b_0(2-\lambda_1/b_0)(1+2\varepsilon b_0)$, so $r\lambda < -b_0$ since $|\lambda_1|/b_0 < 1$ and $2\varepsilon b_0 > 0$. This shows that the vector field points in on the outer boundary of A_ε.

(2) $r = 1-2\varepsilon b_0$, we have $r\lambda = b_0(2+\lambda_1/b_0)(1-2\varepsilon b_0)$. But $2+\lambda_1/b_0 \geqq 2-|\lambda_1|/b_0$, so the bounds for $|\lambda_1|/b_0$ and ε show that $2+\lambda_1/b_0 > 1-k+\sqrt{k^2+1}$ and $1-2\varepsilon b_0 > (k-1+\sqrt{k^2+1})/(2k)$. Thus, $r\lambda > b_0[\sqrt{k^2+1}-(k-1)]\cdot[\sqrt{k^2+1}+(k-1)]/(2k) = b_0$, so the vector field also points in on the inner boundary of A_ε. □

From now on, we only consider values of ε that are small enough to ensure that A_ε is contained in A and is attracting.

3. The estimates in A_ε. In this section, basic estimates are proved for the solutions of (2.2) in A_ε. In particular, it will be shown that the angular component can be approximated by the solution of an equation of the form $\dot{\psi} = 1 - bp(\sigma t)\sin\psi$, which will be analyzed in the next section. As will be seen in § 5, shear has a significant effect on entrainment when it is of the same order of magnitude as the strength of attraction of the limit cycle. In order to allow for this, it will from now on be assumed that β depends on ε in such a way that $\varepsilon/\beta \to c$ as $\varepsilon \to 0$, where c can be any real number.

We begin by rewriting (2.2) as

$$\begin{aligned} \dot{r} &= -\frac{1}{\varepsilon}(r-1)r + r\lambda_1 + b_0 p(\sigma t)\cos\eta, \\ \dot{\eta} &= 1 + \frac{1}{\beta}(r-1) + \omega_1 - (b_0/r)p(\sigma t)\sin\eta. \end{aligned} \tag{3.1}$$

The terms $r\lambda_1$ and ω_1 are $O(\varepsilon)$ in A_ε. Indeed, since $|r-1| \leqq 2b_0\varepsilon$ in A_ε, an immediate consequence of Lemma 2.1 is the following.

LEMMA 3.1. *The following hold in A_ε:*

(i) $|\lambda_1(r,\varepsilon)| \leqq 4b_0^2 k\varepsilon$;

(ii) $|\lambda_{1r}(r,\varepsilon)| \leqq 2b_0 k$;

(iii) $|\omega_1(r,\beta)| \leqq 4b_0^2 k|\gamma|\varepsilon$;

(iv) $|\omega_{1r}(r,\beta)| \leqq 2b_0 k|\gamma|$,

where $\gamma(\varepsilon) = \varepsilon/\beta \to c$ as $\varepsilon \to 0$. □

When $c = 0$, the term $(r-1)/\beta \to 0$ as $\varepsilon \to 0$. In that case, the equation for η can be approximated by $\dot{\eta} = 1 - b_0 p(\sigma t)\sin\eta$. However, when $c \neq 0$, $(r-1)/\beta$ is only bounded, so such an approximation would not be justified. In order to handle this case as well, let $\gamma(\varepsilon)$ be as in Lemma 3.1, let $b(\varepsilon) = b_0\sqrt{1+\gamma^2}$, $d(\varepsilon) = \tan^{-1}\gamma$ and make the change of variable $\theta = \eta + \gamma \ln r - d$ to get

$$\dot{r} = -\frac{1}{\varepsilon}(r-1)r + r\lambda_1 + b_0 p(\sigma t)\cos(\theta - \gamma\ln r + d), \tag{3.2a}$$

$$\dot{\theta} = 1 + \omega_1 + \gamma\lambda_1 - \left(\frac{b}{r}\right)p(\sigma t)\sin(\theta - \gamma\ln r), \tag{3.2b}$$

eliminating the $(r-1)/\beta$ term.

Now let $\psi(t; \psi_0)$ denote the general solution of

$$\dot{\psi} = 1 - b_1 p(\sigma t) \sin \psi, \qquad \psi(0; \psi_0) = \psi_0, \tag{3.3}$$

where $b_1 = \lim_{\varepsilon \to 0} b(\varepsilon) = b_0\sqrt{1+c^2}$, and let $r(t; r_0, \theta_0)$ and $\theta(t; r_0, \theta_0)$ denote the general solution of (3.2), with $r(0; r_0, \theta_0) = r_0$ and $\theta(0; r_0, \theta_0) = \theta_0$.

Standard theorems ensure that these solutions exist for all $t \geqq 0$ when $r_0 \in I_\varepsilon$, and have continuous first partial derivatives with respect to their arguments [6]. Let Ψ, R_i, Θ_i $(i = 1, 2)$ be given by

$$\Psi = \frac{\partial}{\partial \psi_0} \psi(t; \psi_0),$$

$$R_1 = \frac{\partial}{\partial r_0} r(t; r_0, \theta_0), \qquad R_2 = \frac{\partial}{\partial \theta_0} r(t; r_0, \theta_0),$$

$$\Theta_1 = \frac{\partial}{\partial r_0} \theta(t; r_0, \theta_0), \qquad \Theta_2 = \frac{\partial}{\partial \theta_0} \theta(t; r_0, \theta_0).$$

Observe that the periodicity of (3.2) in θ and (3.3) in ψ implies that r, Θ_i, and R_i are 2π-periodic in θ_0, and Ψ is 2π-periodic in ψ_0, with $\theta(t; r_0, \theta_0 + 2\pi) = \theta(t; r_0, \theta_0) + 2\pi$ and $\psi(t; \psi_0 + 2\pi) = \psi(t; \psi_0) + 2\pi$.

In the following lemma and theorem, which are the main results of this section, T is a positive number and $S_{T,\varepsilon}$ denotes the set $[0, T] \times I_\varepsilon \times \mathbb{R}$.

LEMMA 3.2. *There are positive constants $\varepsilon_1(b_0, T)$, $k_1(b_0, T)$, and $k_2(b_0, T)$ such that*

(i) $|R_1(t; r_0, \theta_0)| < e^{-t/2\varepsilon} + \varepsilon$,

(ii) $|\Theta_1(t; r_0, \theta_0)| < \varepsilon k_1$,

(iii) $|R_2(t; r_0, \theta_0)| < 2\varepsilon b_0 k_2$, *and*

(iv) $|\Theta_2(t; r_0, \theta_0)| < k_2$

for all $(t, r_0, \theta_0) \in S_{T,\varepsilon}$ whenever $\varepsilon < \varepsilon_1$.

THEOREM 3.3. *For every $\delta > 0$ there is an $\varepsilon_2(b_0, T) > 0$ such that*

(i) $|\theta(t; r_0, \theta_0) - \psi(t; \theta_0)| < \delta$, *and*

(ii) $|\Theta_2(t; r_0, \theta_0) - \Psi(t; \theta_0)| < \delta$

for all $(t, r_0, \theta_0) \in S_{T,\varepsilon}$ whenever $\varepsilon < \varepsilon_2$.

In addition to the explicit dependence on b_0 and T, these estimates also depend on c, k and the functions λ, ω and $\beta(\varepsilon)$ generally. On the other hand, it will be seen that they are independent of σ since $|p(\sigma t)| \leqq 1$.

Since the variational equations for (3.2) and (3.3) will be used in the proof of Lemma 3.2 and Theorem 3.3, we begin by observing that the functions R_i, Θ_i and Ψ satisfy

$$\dot{\Psi} = -(b_1 p(\sigma t) \cos \psi)\Psi, \tag{3.4}$$

$$\dot{R}_i = B_1 R_i + B_2 \Theta_i, \tag{3.5a}$$

$$\dot{\Theta}_i = B_3 R_i + B_4 \Theta_i, \tag{3.5b}$$

with $R_1(0)=\Theta_2(0)=\Psi(0)=1$ and $R_2(0)=\Theta_1(0)=0$, where the coefficients of (3.5) are given by

$$B_1=-\frac{1}{\varepsilon}(2r-1)+\lambda_1+r\lambda_{1r}+\left(\frac{\gamma b_0}{r}\right)p(\sigma t)\sin(\theta-\gamma\ln r+d),$$

$$B_2=-b_0p(\sigma t)\sin(\theta-\gamma\ln r+d),$$

$$B_3=\omega_{1r}+\gamma\lambda_{1r}+\left(\frac{b}{r^2}\right)p(\sigma t)[\sin(\theta-\gamma\ln r)+\gamma\cos(\theta-\gamma\ln r)],$$

$$B_4=-\left(\frac{b}{r}\right)p(\sigma t)\cos(\theta-\gamma\ln r).$$

Estimates for these coefficients follow easily from Lemma 3.1, together with the fact that $r\to 1$, $\gamma\to c$ and $b\to b_1$ as $\varepsilon\to 0$, yielding the following lemma.

LEMMA 3.4. *There is an* $M(b_0, c, k)>0$ *such that*

(i) $B_1\leqq -\dfrac{1}{2\varepsilon}$,

(ii) $|B_2|\leqq b_0$,

(iii) $|B_3|\leqq M$,

(iv) $|B_4|\leqq 2b_1$,

for $r\in I_\varepsilon$ *whenever* ε *is small enough.* □

In addition to Lemma 3.4, the following two lemmas will be needed. Lemma 3.5 plays the same role as the Gronwall lemma, but seems more convenient here. The proofs are omitted, since Lemma 3.5 is a special case of a lemma of Kelley [7] and the proof of Lemma 3.6 is elementary.

LEMMA 3.5. *Let* T, a, b *be positive constants, and let* v *be defined and continuous on* $[0, T]$, *smooth on* $(0, T)$, *with* $v(0)=0$ *and* $|\dot v|<a|v|+b$ *for* $t\in(0, T)$. *Then* $|v|<(b/a)(e^{at}-1)$ *for* $t\in(0, T]$. □

LEMMA 3.6. *Let* $T>0$ *be given, and let* f, g, h, i, u, v *be real valued and continuous on* $[0, T]$, u, v *smooth on* $(0, T)$, *with* $\dot v=f(t)+g(t)v$, $\dot u=h(t)+i(t)u$ *and* $u(0)\geqq|v(0)|$. *If* $h(t)>|f(t)|$ *and* $i(t)\geqq g(t)$ *for* $t\in(0, T)$, *then* $|v(t)|<u(t)$ *for* $t\in(0, T]$. □

We can now prove Lemma 3.2 and Theorem 3.3. In doing so, the fact that $r(t; r_0, \theta_0)$ is in I_ε for all $t\geqq 0$ if $r_0\in I_\varepsilon$ will be used without explicit mention.

Proof of Lemma 3.2. Let M be as in Lemma 3.4, and let

$$k_1=M\left(2+\frac{1}{2b_0}\right)e^{2b_1T},\qquad k_2=2e^{2b_1T}.$$

(i) and (ii). Consider only $\varepsilon<1/(2b_0k_1)$, and assume one of these is false. Then since they hold for $t=0$, there is a $\tau\in(0, T]$ such that they hold on $[0, \tau)$ but at least one of them fails at $t=\tau$. Then from (3.5a) and Lemma 3.4 we see that

$$\dot R_1=f_1(t)+g_1(t)R_1,\qquad R_1(0)=1,$$

where $|f_1(t)|<\frac{1}{2}$ and $g_1(t)\leqq -1/2\varepsilon$ on $(0, \tau)$. Lemma 3.6 shows that $|R_1|<u_1(t)$ on $(0, \tau]$, where

$$\dot u_1=\frac{1}{2}-\frac{1}{2\varepsilon}u_1,\qquad u_1(0)=1.$$

But $u_1 = (1-\varepsilon)e^{-t/2\varepsilon} + \varepsilon$, so $|R_1| < e^{-t/2\varepsilon} + \varepsilon$ on $[0, \tau]$, showing that (i) cannot have failed. Together with (3.5b) and Lemma 3.4, this shows that

$$\dot{\Theta}_1 = f_2(t) + g_2(t)\Theta_1, \qquad \Theta_1(0) = 0,$$

where $|f_2(t)| < M(e^{-t/2\varepsilon} + \varepsilon)$ and $g_2(t) \leqq 2b_1$ on $(0, \tau)$. Now Lemma 3.6 yields $|\Theta_1| < u_2(t)$ on $(0, \tau]$, where

$$\dot{u}_2 = M(e^{-t/2\varepsilon} + \varepsilon) + 2b_1 u_2, \qquad u_2(0) = 0.$$

But

$$u_2 = \varepsilon M\left[\left(\frac{2}{1+4\varepsilon b_1} + \frac{1}{2b_1}\right) e^{2b_1 t} - \left(\frac{2}{1+4\varepsilon b_1} e^{-t/2\varepsilon} + \frac{1}{2b_1}\right)\right],$$

so we get

$$|\Theta_1| < \varepsilon M\left(\frac{2}{1+4\varepsilon b_1} + \frac{1}{2b_1}\right) e^{2b_1 t} < \varepsilon M\left(2 + \frac{1}{2b_0}\right) e^{2b_1 T} = \varepsilon k_1$$

on $[0, \tau]$, having used $b_1 \geqq b_0 > 0$. But then (ii) cannot have failed at $t = \tau$ either, a contradiction. This establishes (i) and (ii).

(iii) and (iv). Now consider only $\varepsilon < 1/(Mk_2)$, and assume one of these is false. The proof in this case proceeds exactly as above, except now we have $|f_1(t)| < b_0 k_2$,

$$|f_2(t)| < 2\varepsilon b_0 k_2 M, \quad u_1(0) = 0, \quad u_2(0) = 1,$$

yielding

$$u_1 = 2\varepsilon b_0 k_2 (1 - e^{-t/2\varepsilon}), \qquad u_2 = e^{2b_1 t} + \varepsilon k_2 M \frac{b_0}{b_1}(e^{2b_1 t} - 1),$$

so then $|R_2| < 2\varepsilon b_0 k_2$ and

$$|\Theta_2| < (1 + \varepsilon k_2 M)\, e^{2b_1 t} < 2e^{2b_1 T} = k_2 \quad \text{on } [0, \tau]. \qquad \square$$

Proof of Theorem 3.3. (i) From (3.2b) and (3.3), it follows that

$$|\dot{\theta}(t; r_0, \theta_0) - \dot{\psi}(t; \theta_0)| = |b_1 p(\sigma t)(\sin \psi - \sin \theta) + \Delta|,$$

where $\Delta = \omega_1 + \gamma\lambda_1 + p(\sigma t)[b_1 \sin \theta - (b/r) \sin(\theta - \gamma \ln r)]$. Since $|p(\sigma t)| \leqq 1$, this shows that

$$|\dot{\theta} - \dot{\psi}| \leqq b_1 |\theta - \psi| + |\Delta|.$$

Now (i) and (ii) of Lemma 3.1, together with the fact that $\gamma \to c$, $b \to b_1$ and $r \to 1$ as $\varepsilon \to 0$, shows that $|\Delta|$ can be made less than any $\delta_1 > 0$ by taking ε small enough (depending on b_0 and the functions λ, ω and $\beta(\varepsilon)$). Consequently, Lemma 3.5 allows us to conclude that

$$|\theta(t; r_0, \theta_0) - \psi(t; \theta_0)| < \frac{\delta_1}{b_1}(e^{b_1 T} - 1)$$

for $(t, r_0, \theta_0) \in S_{T,\varepsilon}$.

(ii) From (3.4) and (3.5b) we get

$$|\dot{\Theta}_2 - \dot{\Psi}| \leqq |(b_1 \cos \psi)\Psi - (b/r)\Theta_2 \cos(\theta - \gamma \ln r)| + |B_3 R_2|,$$

where $\psi = \psi(t; \theta_0)$ and we have used $|p(\sigma t)| \leqq 1$. Adding and subtracting $\Theta_2 \cos \psi$ in the first term on the right, and using Lemmas 3.2 and 3.4, we find that

$$|\dot{\Theta}_2 - \dot{\Psi}| < b_1 |\Theta_2 - \Psi| + k_2(|b_1 \cos \psi - (b/r) \cos(\theta - \gamma \ln r)| + 2\varepsilon b_0 M)$$

on $S_{T,\varepsilon}$. Then (i) above, together with the fact that $r \to 1$, $b \to b_1$, and $\gamma \to c$ as $\varepsilon \to 0$, shows that

$$|\dot{\Theta}_2 - \dot{\Psi}| < b_1 |\Theta_2 - \Psi| + \delta_1$$

on $S_{T,\varepsilon}$ for any $\delta_1 > 0$ whenever ε is small enough. Since $(\Theta_2 - \Psi)(0) = 0$, Lemma 3.5 yields

$$|\Theta_2 - \Psi| \leqq \frac{\delta_1}{b_1}(e^{b_1 t} - 1) \leqq \frac{\delta_1}{b_1}(e^{b_1 T} - 1) \quad \text{on } S_{T,\varepsilon}. \qquad \square$$

4. The phase approximation. The solution ψ of (3.3) is analyzed in this section, with the intention of discovering when there are values of u for which $\psi(t; u)$ is periodic (mod 2π). The main result that will be obtained asserts that whenever b_1 is large enough (depending on p and σ) there will be exactly two values of u in $[-\pi, \pi)$ for which $\psi(t + 2\pi/\sigma; u) = \psi(t; u) + 2\pi$, with one corresponding to a sink and the other to a source, and all other values in $[-\pi, \pi)$ corresponding to solutions that approach the sink (mod 2π). Together with the results of § 3, this will be used in the next section to obtain results concerning entrainment in (2.1).

4.1. Analysis of $\psi(t; u)$. We begin by writing (3.3) as

$$\dot{\psi} = 1 - bp(\sigma t) \sin \psi, \tag{4.1}$$

where for this section only we write b for b_1. Thus, b here represents a constant, and not the function $b(\varepsilon)$ that it denotes in the rest of the paper. As is easily seen, any periodic solution of (4.1) must have an integer multiple of $2\pi/\sigma$ as its fundamental period. A useful tool in discussing such solutions is the rotation number ρ, given here by

$$\rho = \frac{1}{\sigma} \lim_{t \to \infty} \frac{\psi(t; u)}{t}.$$

As is well known, ρ exists independently of u, depends continuously on parameters, and is such that (4.1) will have a periodic solution of fundamental period $2N\pi/\sigma$ for which $\psi(t + 2N\pi/\sigma) = \psi(t) + 2M\pi$ if and only if ρ is rational, with $\rho = M/N$ in lowest terms [8]. An easy result is that $\rho > 0$ for (4.1), ruling out solutions that are periodic in the ordinary sense. To see this, let $f(t) = (\psi(t) - \psi(-t))/2$, where ψ is any solution of (4.1), and observe that $\dot{f} = 1$ whenever $f = 0$. Thus, $t = 0$ is the only zero of f and $f > 0$ for $t > 0$. Since (iv) of Lemma 4.1 (below) shows that ψ can be replaced by f in the expression for ρ, this shows that $\rho \geqq 0$. But $\rho = 0$ cannot occur, since if ψ was periodic f would have infinitely many zeros.

We now prove two lemmas that will provide the basic results needed to analyze the Poincaré map $u: \mapsto \psi(2\pi/\sigma; u)$ (mod 2π). From the symmetry properties of $p(\sigma t)$ and $\sin \psi$, we get

LEMMA 4.1. *The following hold for all real values of t and u:*

(i) $\psi(t; u) + 2\pi = \psi(t; u + 2\pi)$;

(ii) $\psi\left(t + \frac{2\pi}{\sigma}; u\right) = \psi\left(t; \psi\left(\frac{2\pi}{\sigma}; u\right)\right)$;

(iii) $\psi\left(t + \frac{\pi}{\sigma}; u\right) - \pi = \psi\left(t; \psi\left(\frac{\pi}{\sigma}; u\right) - \pi\right)$;

(iv) $-\psi(-t; u) = \psi(t; -u)$.

Proof. In each case, the function on the left is a solution of (4.1), so the result follows from uniqueness. □

One consequence of this lemma is that we need only consider $t \in [0, \pi/\sigma]$. The next lemma provides estimates for $\psi(t; 0)$ and $\psi(t; \pi)$ on this interval. In order to state it, we first define positive numbers $m(a_1, a_2)$ and $\bar{b}(\sigma, \delta)$ for $[a_1, a_2] \subset (0, \pi)$ and $0 < \delta \leqq \min(\pi/2, \pi/2\sigma)$ by

$$m(a_1, a_2) = \min_{s \in [a_1, a_2]} p(s),$$

$$\bar{b}(\sigma, \delta) = \max\,(1/[m(\sigma\delta, \pi - \sigma\delta) \sin \delta], (2\pi - 4\delta)/[\delta m(\sigma\delta/2, \sigma\delta) \sin (\delta/2)]).$$

Then we have the following lemma.

LEMMA 4.2. *If* $0 < \delta \leqq \min(\pi, \pi/\sigma)$, *then*

(i) $\psi(t; 0) < \delta$ *for* $t \in [0, \delta]$;

(ii) $\psi(t; -\pi) > -\pi + \delta$ *for* $t \in \left[\delta, \dfrac{\pi}{\sigma}\right]$.

If, in addition, $\delta \leqq \min(\pi/2, \pi/2\sigma)$ *and* $b > \bar{b}(\sigma, \delta)$, *then*

(iii) $\psi(t; 0) < \delta$ *for* $t \in \left[0, \dfrac{\pi}{\sigma} - \delta\right]$, *with*

$\psi(t; 0) < 2\delta$ *for* $t \in \left(\dfrac{\pi}{\sigma} - \delta, \dfrac{\pi}{\sigma}\right]$;

(iv) $\psi(t; -\pi) > -\delta$ *for* $t \in [\delta, 2\delta)$, *with*

$\psi(t; -\pi) > 0$ *for* $t \in \left[2\delta, \dfrac{\pi}{\sigma}\right]$.

Proof. (i) Since $\dot{\psi} < 1$ for $(t, \psi) \in (0, \delta] \times (0, \delta]$ it follows immediately that $\psi(t; 0) < \delta$ for $t \in [0, \delta]$ since $\psi(0; 0) = 0$.

(ii) Similarly, $\dot{\psi} > 1$ for $(t, \psi) \in (0, \delta] \times (-\pi, -\pi + \delta]$, so $\psi(\delta; -\pi) > -\pi + \delta$ since $\psi(0; -\pi) = -\pi$. But $\dot{\psi} \geqq 1$ at $\psi = -\pi + \delta$ for $t \in [0, \pi/\sigma]$ so $\psi(t; -\pi)$ cannot return to $-\pi + \delta$ until after $t = \pi/\sigma$.

(iii) We have $\psi(t; 0) < \delta$ for $t \in [0, \delta]$ by (i). But $b > \bar{b}$ implies that $\dot{\psi} < 0$ at $\psi = \delta$ for $t \in [\delta, \pi/\sigma - \delta]$, since then $\dot{\psi} = 1 - bp(\sigma t) \sin \delta$ is less than $1 - p(\sigma t)/m(\sigma\delta, \pi - \sigma\delta) \leqq 0$. Thus $\psi(t; 0) < \delta$ for $t \in [0, \pi/\sigma - \delta]$. To prove the rest, observe that $\dot{\psi} \leqq 1$ for $(t, \psi) \in [\pi/\sigma - \delta, \pi/\sigma] \times [\delta, \pi]$ and that $\pi - \delta \geqq \delta$ since $\delta \leqq \pi/2$. Together with $\psi(\pi/\sigma - \delta; 0) < \delta$, this shows that $\psi(t; 0) < 2\delta$ for $t \in [\pi/\sigma - \delta, \pi/\sigma]$.

(iv) Since $\dot{\psi} = 1$ at $\psi = 0$ for all t, and since $\dot{\psi} > 1$ for $(t, \psi) \in (0, 2\delta] \times [-\delta, 0)$, all that is needed is to show that $\psi(t; -\pi)$ reaches $-\delta$ before $t = \delta$. From (ii) we get $\psi(\delta/2; -\pi) > -\pi + \delta/2$. Thus it is sufficient to show that $(\delta/2)\dot{\psi} > \pi - 3\delta/2$ for $b > \bar{b}$ and $(t, \psi) \in [\delta/2, \delta] \times [-\pi + \delta/2, -\delta]$. But for such t, ψ and b we have $-bp(\sigma t) \sin \psi > (2\pi - 4\delta)/\delta$ yielding

$$\frac{\delta}{2}\dot{\psi} > \frac{\delta}{2}\left[1 + \frac{2\pi - 4\delta}{\delta}\right] = \pi - \frac{3\delta}{2}.$$ □

4.2. The Poincaré map and its square root. In order to discuss periodic solutions of (4.1), we now introduce the Poincaré map $h(u) = \psi(2\pi/\sigma; u) - 2\pi$ and the related function $g(u) = \psi(\pi/\sigma; u) - \pi$. Both of these functions are increasing diffeomorphisms of $\mathbb{R}$. Furthermore, since the n-fold composition $h^n(u)$ equals $\psi(2n\pi/\sigma; u) - 2n\pi$ by

(ii) of Lemma 4.1, $\psi(t; u)$ will be a $(2n\pi/\sigma)$-periodic solution of (4.1) if and only if $h^n(u) = u \pmod{2\pi}$. In particular, fixed points of h correspond to $(2\pi/\sigma)$-periodic solutions with $\rho = 1$. From Lemma 4.1 we get the following.

LEMMA 4.3. *g and h have the following properties*:

(i) $g(u+2\pi) = g(u)+2\pi, \qquad h(u+2\pi) = h(u)+2\pi;$

(ii) $h = g \circ g;$

(iii) $g^{-1}(u) = -g(-u).$

Proof. (i) follows immediately from (i) of Lemma 4.1. Similarly, (ii) follows from (iii) of Lemma 4.1 with $t = \pi/\sigma$. To prove (iii), we have

$$g(-g(-u)) = \psi\left(\frac{\pi}{\sigma}; -\psi\left(\frac{\pi}{\sigma}; -u\right) + \pi\right) - \pi,$$

which equals

$$-\psi\left(-\frac{\pi}{\sigma}; \psi\left(\frac{\pi}{\sigma}; -u\right) - \pi\right) - \pi$$

by (iv) of Lemma 4.1. Now (iii) of Lemma 4.1 shows that $g(-g(-u)) = -\psi(0; -u) = u$. □

As a result of (ii) above, h can be analyzed in terms of its "square root" g. This simplifies the analysis, since $p(\sigma t)$ is of constant sign on $[0, \pi/\sigma]$ and because it allows easy use of Lemma 4.2. The next result provides a sufficient condition for $\rho = 1$.

LEMMA 4.4. *Let* $0 < \delta \leqq \min(\pi/2, \pi/2\sigma)$. *If* $b > \bar{b}(\sigma, \delta)$, *then* $g([-\pi, 0]) \subset (-\pi, -\pi+2\delta)$ *and* $g^{-1}([0, \pi]) \subset (\pi-2\delta, \pi)$. *If* $\sigma = 1$, *then* $g([-\pi, 0]) \subset (-\pi, 0)$ *and* $g^{-1}([0, \pi]) \subset (0, \pi)$ *for all* $b > 0$.

Proof. The result for g^{-1} follows from the one for g by (iii) of Lemma 4.3. To prove the result for g, it is sufficient to prove that

(A) $$\psi\left(\frac{\pi}{\sigma}; 0\right) < \pi \quad \text{and} \quad \psi\left(\frac{\pi}{\sigma}; -\pi\right) > 0 \quad \text{for } \sigma = 1, \quad \text{and}$$

(B) $$\psi\left(\frac{\pi}{\sigma}; 0\right) < 2\delta \quad \text{and} \quad \psi\left(\frac{\pi}{\sigma}; -\pi\right) > 0 \quad \text{for } b > \bar{b}(\sigma, \delta).$$

But (A) follows from (i) and (ii) of Lemma 4.2 with $\delta = \pi$ and $t = \pi/\sigma$, while (B) follows from (iii) and (iv) of Lemma 4.2 with $t = \pi/\sigma$. □

Since $(-\pi, -\pi+2\delta) \subset (-\pi, 0)$ for δ as above, g clearly has fixed points when the conclusions of the lemma hold. Furthermore, fixed points of g are also fixed points of h since $h = g \circ g$, so Lemma 4.4 provides sufficient conditions for (4.1) to have rotation number 1. In particular, observe that $\rho = 1$ for all $b > 0$ when $\sigma = 1$. This cannot happen for any other value of σ since $\rho \to 1/\sigma$ as $b \to 0$, ruling out $\rho = 1$ for b small when $\sigma \neq 1$.

We now consider the questions of the number of fixed points in $[-\pi, \pi)$ and their stability. These questions can be answered by examining g', with a fixed point u of g being stable if $g'(u) < 1$ and unstable if $g'(u) > 1$. To begin, since $g'(u) = \Psi(\pi/\sigma; u)$ we can solve (3.4) to get

(4.2) $$g'(u) = \exp\left(-b \int_0^{\pi/\sigma} p(\sigma t) \cos \psi(t; u)\, dt\right).$$

Since $p(s)$ has maximum value 1 on $(0, \pi)$, there are numbers $0 < a_1 < a_2 < \pi$ for which $p(s) \geqq 1/\sqrt{2}$ on $[a_1, a_2]$. Now let the positive numbers δ and ν be given by

$$\nu = \frac{a_2 - a_1}{4}, \qquad \delta = \min\left(\frac{\pi}{4}, \frac{\pi}{2\sigma}, \frac{a_1}{\sigma}, \frac{\pi - a_2}{\sigma}, \frac{\nu}{\sigma}\right).$$

We can now prove the following.

THEOREM 4.5. *Let ν and δ be as above. Then*

(i) *If $b > \bar{b}(\sigma, \delta)$, then g is a contraction mapping on $[-\pi, 0]$, with $g' < \exp(-b\nu/\sigma)$ there.*

(ii) *If $\sigma = 1$, then there is an interval $J \subset (-\pi, 0)$ (depending on p) such that for all sufficiently small b, g is a contraction mapping on J, with $g' < \exp(-b\nu_2)$ there, and all points of $[-\pi, 0]$ not in J enter it under iteration of g. Here ν_2 is a positive constant determined by p.*

Furthermore, the corresponding assertions hold for g^{-1} on $[0, \pi]$ and $-J = \{x \mid -x \in J\}$.

Proof. Since $g^{-1}(u) = -g(-u)$ (Lemma 4.3), the assertions for g^{-1} follow from those for g. The proofs for g are as follows.

(i) For $b > \bar{b}(\sigma, \delta)$, we have $g([-\pi, 0]) \subset (-\pi, 0)$ by Lemma 4.4. To get the rest, we only need to show that the integral in (4.2) will be greater than ν/σ for $u \in [-\pi, 0]$. But for such u, it follows from (iii) and (iv) of Lemma 4.2 that $\psi(t; u) \in (-\delta, 2\delta)$ for $t \in [\delta, \pi/\sigma]$, with $\psi(t; u) \in (-\delta, \delta)$ for $t \in [\delta, \pi/\sigma - \delta]$. Furthermore, we have $(-\delta, 2\delta) \subset (-\pi/4, \pi/2)$, $(-\delta, \delta) \subset (-\pi/4, \pi/4)$, and $p(\sigma t) \geqq 1/\sqrt{2}$ for $t \in [a_1/\sigma, a_2/\sigma] \subset [\delta, \pi/\sigma - \delta]$. Thus, the negative contribution to the integral must be greater than $-\delta$, with $p(\sigma t) \cos \psi(t; u) > \frac{1}{2}$ for $t \in [a_1/\sigma, a_2/\sigma]$, so

$$\int_0^{\pi/\sigma} p(\sigma t) \cos \psi(t; u)\, dt > \frac{a_2 - a_1}{2\sigma} - \delta \geqq \frac{a_2 - a_1}{2\sigma} - \frac{\nu}{\sigma} = \frac{\nu}{\sigma}.$$

(ii) Now let $\sigma = 1$. Expanding $g(u)$ and $\cos \psi(t; u)$ in powers of b and making use of (4.1) and (4.2), we get expansions that can be written in the form

$$(*) \qquad g(u) = u - bA \sin(u + \xi) + O(b^2),$$

$$(**) \qquad g'(u) = \exp(-bA \cos(u + \xi) + O(b^2))$$

as $b \to 0$, uniformly in (t, u) for $t \in [0, \pi]$. Here $A = (\alpha^2 + \beta^2)^{1/2}$ and $\xi = \operatorname{arc\,cot}(\beta/\alpha)$, where

$$\alpha = \int_0^{\pi} p(t) \sin t\, dt \quad \text{and} \quad \beta = \int_0^{\pi} p(t) \cos t\, dt.$$

Since $\xi \in (0, \pi)$, there is a closed interval $J \subset (-\pi, 0)$, containing $-\xi$ in its interior, on which $\cos(u + \xi) > \frac{3}{4}$. Clearly $\sin(u + \xi)$ is bounded away from 0 on $[-\pi, 0] - J$, where it is positive to the right of J, and negative to the left. Together with (*), this shows that $g(J) \subset J$ and that points in $[-\pi, 0] - J$ enter J under iteration of g, whenever b is small enough.

To show that g is a contraction on J, we need only establish the estimate for g'. To this end, let $\nu_2 = A/2$. Since $\cos(u + \xi) > \frac{3}{4}$ on J, it follows from (**) that $g'(u) < \exp(-b\nu_2)$ there whenever b is small enough. □

When the conclusions of Theorem 4.5 hold, g (and so also h) has exactly two fixed points in $[-\pi, \pi)$: a sink $u_0 \in (-\pi, 0)$ and a source $-u_0 \in (0, \pi)$. Furthermore, if u is any other point of $[-\pi, \pi)$, then u approaches u_0 under iteration of g if $u < -u_0$ and approaches $u_0 + 2\pi$ if $u > -u_0$. The behavior of the corresponding map $\tilde{g}: S^1 \to S^1$ is sketched in Fig. 1.

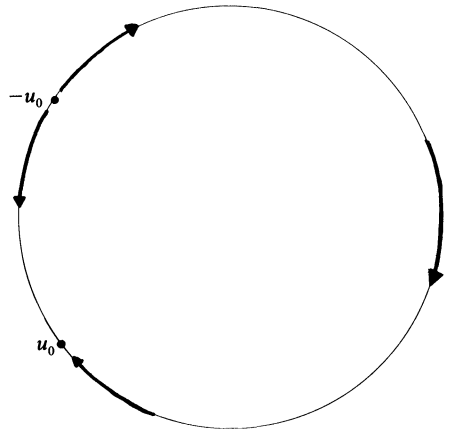


FIG. 1. *Behavior of $\tilde{g}$ when b is large and also for small b when $\sigma = 1$.*

Remark 1. In addition to their existence, we can obtain information about the location of the fixed points when b is large. In particular, $g([-\pi, 0]) \subset (-\pi, -\pi + 2\delta)$ for $b > \bar{b}(\sigma, \delta)$ by Lemma 4.4, which shows that $u_0 \to -\pi$ as $b \to \infty$.

Remark 2. It should be pointed out that it was not obviously to be expected that (4.1) would have rotation number 1 for large b. To see this, consider the equation

$$\dot{\psi} = 1 - bp(\sigma t)(1 + \sin \psi), \tag{4.3}$$

which appears similar to (4.1). If we let $\tau = \sigma t - \pi$ and introduce $v(\tau)$ by

$$\frac{v_\tau}{v} = \frac{-1}{2\sigma} \tan\left(\frac{\psi}{2} - \frac{\pi}{4}\right),$$

then (4.3) is transformed into

$$v_{\tau\tau} + (\alpha + \beta p(\tau))v = 0 \qquad \left(\alpha = \frac{1}{4\sigma^2},\ \beta = \frac{1}{2\sigma^2}\right) \tag{4.4}$$

which is Hill's equation. Using an argument given by Ermentrout [9] for the case $p(\tau) = \sin(\tau)$, and presented more fully by Ermentrout and Kopell [10] in the general case, it can be shown that (4.3) has rotation number k when (α, β) is in the kth *instability* zone of (4.4) ($k = 1, 2, 3, \cdots$).

When $p(\tau) = \sin(\tau)$, (4.4) is Mathieu's equation, for which the well-known stability diagram shows clearly that as b increases it passes in succession through an infinite sequence of intervals $(b_k, \tilde{b}_k)$, with $b_k < \tilde{b}_k < b_{k+1}$ and $b_k \to \infty$ as $k \to \infty$, on which (4.3) has rotation number k, which is in striking contrast to the behavior of (4.1). This can also be shown to occur when $p(\tau)$ is the square wave with $p = 1$ on $[0, \pi)$ and $p = -1$ on $[\pi, 2\pi)$, in which case explicit conditions for the boundaries of the instability zones can be obtained [10], [11]. Finally, work done by Weinstein and Keller [12] suggests that the relevant aspects of the instability zones in these two cases occur also for a large class of functions p.

5. Entrainment of (2.1). The results of the previous sections will now be used to obtain information concerning the behavior of the solutions of (2.1) in A_ε when ε is small. In particular, the fact that the θ-component of the solutions of (3.2) can be approximated by solutions of (4.1) will allow us to conclude that (2.1) will be 1:1 phase-locked for small ε whenever the conclusions of Theorem 4.5 hold. The main

result, which is contained in Theorem 5.4, asserts that if b_1 and σ satisfy the hypotheses of Theorem 4.5, then for all sufficiently small ε the period map of (2.1) will have exactly two fixed points in A_ε, a sink and saddle corresponding to the sink u_0 and source $-u_0$ of g, and that every point of A_ε that is not on the stable manifold of the saddle will be in the basin of attraction of the sink. This will follow naturally from the fact that $(r(\pi/\sigma; r_0, \theta_0), \theta(\pi/\sigma; r_0, \theta_0))$ can be C^1-approximated in $I_\varepsilon \times \mathbb{R}$ by $(1, g(\theta_0))$, as shown by Lemma 3.2 and Theorem 3.3. That (2.1) will then be 1:1 phase-locked is a consequence of the fact that (4.1) has rotation number 1 whenever g has fixed points.

To begin, let $x(t; x_0, y_0)$ and $y(t; x_0, y_0)$ denote the general solution of (2.1) with $x(0; x_0, y_0) = x_0$ and $y(0; x_0, y_0) = y_0$, and let X_t denote the t-advance map

$$X_t: \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \mapsto \begin{pmatrix} x(t; x_0, y_0) \\ y(t; x_0, y_0) \end{pmatrix}.$$

X_t is defined on A_ε for all $t \geqq 0$ since A_ε is compact and attracting and standard theorems ensure that X_t maps A_ε diffeomorphically onto its image. Furthermore, as shown in the next lemma, X_t has periodicity and symmetry properties similar to those of $\psi(t; u)$.

LEMMA 5.1. *The following hold for all* $t \geqq 0$:

(i) $X_{t+2\pi/\sigma} = X_t \circ X_{2\pi/\sigma}$;

(ii) $-X_{t+\pi/\sigma} = X_t \circ (-X_{\pi/\sigma})$.

Proof. In each case, when the function on the left is evaluated at an arbitrary initial point a solution of (2.1) is obtained. The result then follows from uniqueness. □

From (i) it follows that $X_{2n\pi/\sigma} = X^n_{2\pi/\sigma}$ (n a positive integer), while (ii) shows that $X_{2\pi/\sigma} = (-X_{\pi/\sigma}) \circ (-X_{\pi/\sigma})$. Thus, all information concerning periodic solutions must be contained in $-X_{\pi/\sigma}$ and we can once again restrict our analysis to a half period.

Corresponding to the maps X_t are the maps Φ_t given by

$$\Phi_t: \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix} \mapsto \begin{pmatrix} r(t; r_0, \theta_0) \\ \theta(t; r_0, \theta_0) \end{pmatrix},$$

where (r, θ) is the general solution of (3.2). Let $F = F_\varepsilon$ be the map connecting the coordinates (r, θ) and (x, y),

$$F_\varepsilon: \begin{pmatrix} r \\ \theta \end{pmatrix} \mapsto \begin{pmatrix} r\cos(\theta - \gamma \ln r + d) \\ r\sin(\theta - \gamma \ln r + d) \end{pmatrix}.$$

Then X_t and Φ_t are connected by

$$X_t \circ F = F \circ \Phi_t. \tag{5.1}$$

Observe that

$$D\Phi_t = \begin{pmatrix} R_1 & R_2 \\ \Theta_1 & \Theta_2 \end{pmatrix}, \tag{5.2}$$

which shows that

$$\frac{d}{dt} D\Phi_t = B(t) D\Phi_t, \qquad D\Phi_0 = I \tag{5.3}$$

where

$$B = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}$$

is the coefficient matrix of (3.5) and I is the 2×2 identity matrix. We can now prove the next lemma, which shows that Φ_t contracts areas in $I_\varepsilon \times \mathbb{R}$ for $t>0$ if ε is small enough.

LEMMA 5.2. *If ε is small enough, then*

$$0 < \det D\Phi_t \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix} \leqq e^{(2b_1 - 1/2\varepsilon)t}$$

for all $(r_0, \theta_0) \in I_\varepsilon \times \mathbb{R}$ and all $t>0$.

Proof. From (5.3) it follows that

$$\det D\Phi_t = \exp\left(\int_0^t \operatorname{tr} B(s)\, ds\right).$$

But $\operatorname{tr} B(s) = B_1(s) + B_4(s)$, so Lemma 3.4 shows that $\operatorname{tr} B(s) \leqq 2b_1 - 1/2\varepsilon$ for $r_0 \in I_\varepsilon$ whenever ε is small enough. □

Now consider the map

$$G: \quad \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix} \mapsto \Phi_{\pi/\sigma} \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix} - \begin{pmatrix} 0 \\ \pi \end{pmatrix}.$$

From (5.1) it follows that

$$(-X_{\pi/\sigma}) \circ F = F \circ G. \tag{5.4}$$

Clearly

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = F \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix}$$

will be a fixed point of $-X_{\pi/\sigma}$ if $\binom{r_0}{\theta_0}$ is a fixed point of G. Moreover, if $\binom{r_0}{\theta_0}$ is a hyperbolic fixed point of G for which $r_0 \neq 0$, then $\binom{x_0}{y_0}$ will be a fixed point of the same type (sink, saddle or source) for $-X_{\pi/\sigma}$. To see this, observe that $\det DF\binom{r_0}{\theta_0} = r_0$ and that differentiation of (5.4) yields

$$D(-X_{\pi/\sigma}) \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} DF \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix} = DF \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix} DG \begin{pmatrix} r_0 \\ \theta_0 \end{pmatrix}$$

when $\binom{r_0}{\theta_0}$ is a fixed point of G. Thus, $D(-X_{\pi/\sigma})\binom{x_0}{y_0}$ and $DG\binom{r_0}{\theta_0}$ are similar, and so must have the same eigenvalues, when $\binom{r_0}{\theta_0}$ is a fixed point with $r_0 \neq 0$.

The results of the previous section can easily be used to obtain information about G. This is because Lemma 3.2 and Theorem 3.3 show that G and DG can be made as close as desired to

$$\begin{pmatrix} 1 \\ g \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & 0 \\ 0 & g' \end{pmatrix}$$

for $(r_0, \theta_0) \in I_\varepsilon \times \mathbb{R}$ by making ε small enough (depending on b_1, etc.). In particular, part of the proof of Theorem 5.4 will involve showing that G will be a contraction near $\binom{1}{u_0}$ for ε small. In this regard, observe that a smooth map from a compact convex set to itself will be a contraction if at each point the eigenvalues of its linear part are real, distinct and less than one in absolute value, and the corresponding eigenvectors are mutually orthogonal. The next lemma shows that the orthogonality condition can be relaxed.

LEMMA 5.3. *Let D be a compact convex subset of an open set $U \subset \mathbb{R}^n$ $(n \geqq 2)$, and let $f: U \to \mathbb{R}^n$ be a C^1 function that maps D into itself whose linear part $Df(x)$ has distinct real eigenvalues $\lambda_1(x), \cdots, \lambda_n(x)$ in D. Let $\lambda_M = \max |\lambda_i(x)| (x \in D, i = 1, \cdots, n)$, and let $\mathbf{v}_1(x), \cdots, \mathbf{v}_n(x)$ be unit eigenvectors corresponding to $\lambda_1(x), \cdots, \lambda_n(x)$. Then f will be a contraction mapping on D if $\lambda_M < 1$ and*

$$|\mathbf{v}_i \cdot \mathbf{v}_k| < \frac{1}{n-1} \frac{1-\lambda_M^2}{1+\lambda_M^2} \qquad (i \neq k)$$

for all $x \in D$.

Proof. Since $|\mathbf{v}_i(x) \cdot \mathbf{v}_k(x)|$ is continuous and D is compact, there is a $\lambda \in (\lambda_M, 1)$ such that

$$(*) \qquad |\mathbf{v}_i \cdot \mathbf{v}_k| < \frac{1}{n-1} \frac{\lambda^2 - \lambda_M^2}{\lambda^2 + \lambda_M^2}$$

on D. We now show that $\|Df(x)\mathbf{w}\| \leqq \lambda \|\mathbf{w}\|$ on D for all $\mathbf{w} \in \mathbb{R}^n$, where $\| \cdot \|$ denotes the Euclidean norm. To this end, let $x \in D$ and $\mathbf{w} \in \mathbb{R}^n$ be arbitrary, and let $A = Df(x)$, $B = \operatorname{diag}(\lambda_1(x), \cdots, \lambda_n(x))$, $M = \operatorname{col}(\mathbf{v}_1(x), \cdots, \mathbf{v}_n(x))$ and $N = M^T M - I$, where I is the $n \times n$ identity matrix. M is invertible since $\lambda_1, \cdots, \lambda_n$ are distinct, so we can also let $\mathbf{u} = M^{-1}\mathbf{w}$. We want to show that $\lambda^2\|\mathbf{w}\|^2 - \|A\mathbf{w}\|^2 \geqq 0$. But

$$\lambda^2\|\mathbf{w}\|^2 - \|A\mathbf{w}\|^2 = \mathbf{u}^T(\lambda^2 I - B^2)\mathbf{u} + \mathbf{u}^T(\lambda^2 N - BNB)\mathbf{u}.$$

Clearly $\mathbf{u}^T(\lambda^2 I - B^2)\mathbf{u} \geqq (\lambda^2 - \lambda_M^2)\mathbf{u}^T\mathbf{u}$, and $\lambda^2 N - BNB$ is a symmetric matrix whose diagonal elements are 0 and whose other elements are bounded in absolute value by $(\lambda^2 - \lambda_M^2)/(n-1)$ as follows from (*). As a result we see that

$$\lambda^2\|\mathbf{w}\|^2 - \|A\mathbf{w}\|^2 \geqq \frac{\lambda^2 - \lambda_M^2}{n-1} \mathbf{u}_+^T C \mathbf{u}_+,$$

where $\mathbf{u}_+$ is the vector whose components are the absolute value of the corresponding components of $\mathbf{u}$, and the matrix C has its diagonal elements equal to $n-1$ and its remaining elements equal to -1. Thus, it is sufficient to show that C is a nonnegative matrix. But this follows immediately from the fact that $C = (1/n)C^T C$.

Now let x and y be any two points in D. Since D is convex, the segment joining x and y lies in D and we have

$$\begin{aligned} \|f(y) - f(x)\| &= \left\| \int_0^1 Df(x + t(y-x))(y-x)\, dt \right\| \\ &\leqq \int_0^1 \|Df(x + t(y-x))(y-x)\| dt \\ &\leqq \int_0^1 \lambda \|y - x\| dt = \lambda \|y - x\|. \end{aligned}$$

Thus, f is a contraction on D with contraction constant λ. □

Let δ be as in Theorem 4.5. We can now prove the following.

THEOREM 5.4. *If $b_1 > \bar{b}(\sigma, \delta)$, or if $\sigma = 1$ and b_1 is small enough, then $X_{2\pi/\sigma}$ will have exactly two fixed points in A_ε, one a sink and the other a saddle, whenever ε is small enough (depending on b_1 and σ). Furthermore, all points of A_ε not on the stable manifold of the saddle are in the basin of attraction of the sink.*

The following proof uses the notions of horizontal curve and horizontal strip. Briefly, a smooth horizontal curve in this context is a curve whose tangent vectors have

a θ-component that is larger than its r-component. A horizontal strip is a strip whose upper and lower boundaries are nonintersecting horizontal curves. The width of a horizontal strip is the maximum vertical distance between corresponding points on the upper and lower boundaries. For details, see [14].

Proof. If $b_1 > \bar{b}(\sigma, \delta)$, or if $\sigma = 1$ and b_1 is small enough, then it follows from Theorem 4.5 that there are closed intervals $J_1 \subset (-\pi, 0)$ and $J_2 \subset (0, \pi)$, with $u_0 \in \text{Int}(J_1)$ and $-u_0 \in \text{Int}(J_2)$ (where the sink u_0 is the unique fixed point of g in $(-\pi, 0)$ and the source $-u_0$ is the unique fixed point of g in $(0, \pi)$), such that g' is bounded below 1 on J_1 and bounded above 1 on J_2. Let $D_1 = I_\varepsilon \times J_1$ and $D_2 = I_\varepsilon \times J_2$. Since $X_{2\pi/\sigma} = (-X_{\pi/\sigma}) \circ (-X_{\pi/\sigma})$, it follows from (5.4) that it is sufficient to prove that the following hold whenever ε is small enough.

(A) Points of $I_\varepsilon \times [-\pi, \pi)$ not in $D_1 \cup D_2$ enter $D_1 (\text{mod } 2\pi \text{ in } \theta)$ under iteration of G.

(B) G is a contraction mapping on D_1.

(C) G has a unique saddle point in D_2, and any point of D_2 not on its stable manifold leaves D_2 under iteration of G.

Proof of (A), (B), *and* (C).

(A) and (B). Theorems 4.5 and 3.3 show that there must be a $d > 0$ such that those points $\binom{r}{\theta}$ of $I_\varepsilon \times [-\pi, \pi)$ not in $D_1 \cup D_2$ for which $\theta \in [-\pi, -u_0)$ must move at least a distance d toward D_1 under G, while those with $\theta \in (-u_0, \pi)$ must move at least distance d toward the copy $\{\binom{r}{\theta} | \binom{r}{\theta} - \binom{0}{2\pi} \in D_1\}$ of D_1. This proves (A), and shows that $G(D_1) \subset D_1$. It could now be shown directly that G is a contraction on D_1, but instead we observe that G must satisfy the hypotheses of Lemma 5.3 in D_1 for small ε, since g' is bounded below 1 on J_1 while Lemma 3.2 and Theorem 3.3 show that $DG\binom{r}{\theta}$ can be made as close as desired to

$$\begin{pmatrix} 0 & 0 \\ 0 & g'(\theta) \end{pmatrix}$$

uniformly for $(r, \theta) \in I_\varepsilon \times \mathbb{R}$ by taking ε small enough.

(C). The fact that g' is bounded above 1 on J_2, together with the approximation of DG given above, shows that whenever ε is small enough G will map smooth horizontal curves in D_2 to smooth horizontal curves and DG will have eigenvalues λ_1, λ_2 that satisfy $0 < |\lambda_1| < 1 < |\lambda_2|$ at each point of D_2. Clearly any fixed point of G in D_2 will then be a saddle.

Now let $S_0 = D_2$ and define

$$S_k = D_2 \cap G(S_{k-1}) \qquad (k = 1, 2, 3, \cdots).$$

Since G maps smooth horizontal curves in D_2 to smooth horizontal curves, and since points on the ends of D_2 move at least a distance d away from D_2 under G, it follows that the S_k form a nested family of horizontal strips, each of which connects the ends of D_2. Furthermore, area of S_{k+1} is less than $\frac{1}{2}$ area of S_k whenever ε is small enough, since $\det(DG) < \frac{1}{2}$ for small ε by Lemma 5.2. As a result, the width of $S_k \to 0$ as $k \to \infty$, due to the fact that the upper and lower boundaries of S_k are horizontal curves of length not less than the length of J_2. Thus

$$S_\infty = \bigcap_{k=0}^{\infty} S_k$$

is a horizontal curve [14, p. 70] connecting the ends of D_2 (see Fig. 2).

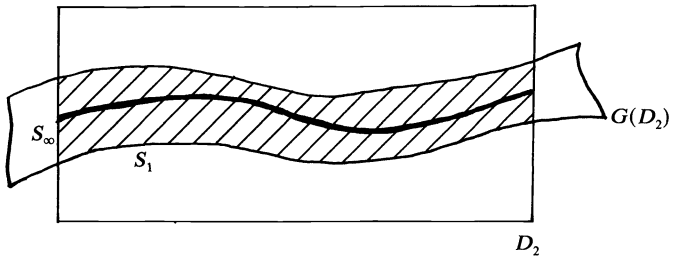


FIG. 2

Now let W^u be the unstable manifold of a fixed point in D_2. Since $D_2 \cap G(D_2 \cap W^u) = D_2 \cap W^u$, it follows that $D_2 \cap W^u \subset S_\infty$. Thus, there can be at most one fixed point in D_2 since otherwise the segment of S_∞ connecting any two fixed points would have to lie on the unstable manifold of each. On the other hand, there must be at least one fixed point in D_2, since $D_2 \cap G(S_\infty) = S_\infty$ shows that $G^{-1}(S_\infty) \subset S_\infty$. We must now conclude that $S_\infty = W^u \cap D_2$. That any points of D_2 not on the stable manifold of the saddle must leave D_2 under iteration of G is a consequence of the fact that such a point must otherwise approach W^u. □

The behavior of $X_{2\pi/\sigma}$ in A_ε when the conclusion of the theorem holds is sketched in Fig. 3. It should be observed that the sink and saddle mentioned in the theorem must approach

$$\begin{pmatrix} \cos(u_0 + d_0) \\ \sin(u_0 + d_0) \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \cos(-u_0 + d_0) \\ \sin(-u_0 + d_0) \end{pmatrix}$$

as $\varepsilon \to 0$, where $d_0 = \tan^{-1} c$. This follows from the form of F, together with the fact that the intervals J_1 and J_2 used in the proof can be chosen as small as desired provided that $u_0 \in \text{Int}(J_1)$ and $-u_0 \in \text{Int}(J_2)$.

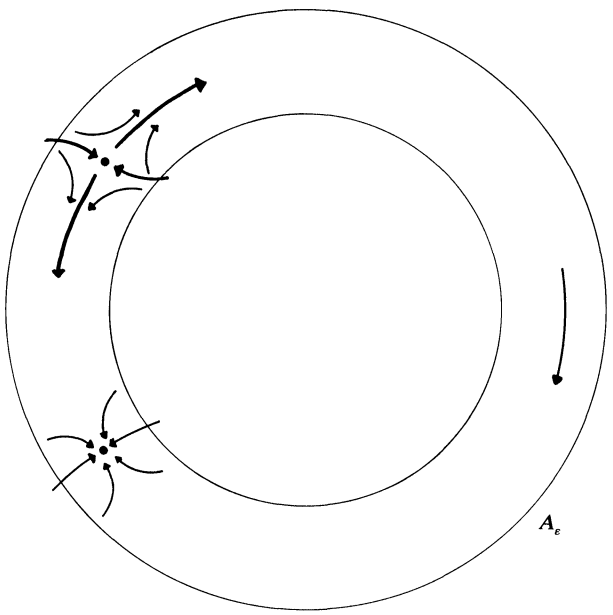


FIG. 3. *Behavior of $X_{2\pi/\sigma}$ in A_ε for small ε when b_1 is large and also for small b_1 when $\sigma = 1$. The width of A_ε has been exaggerated for purposes of illustration.*

Furthermore, since the proof involved approximating the θ-component of solutions of (2.1) by solutions of (4.1) with rotation number 1, it follows that the periodic solutions corresponding to the fixed points of $X_{2\pi/\sigma}$ go once around A_ε in each period, which shows that (2.1) is then 1:1 phase locked. Thus, (2.1) is 1:1 phase locked for small ε when $b_1 > \bar{b}(\sigma, \delta)$, and for b_1 small enough when $\sigma = 1$. That this cannot occur for b_1 small when $\sigma \neq 1$ follows easily from the fact that (4.1) cannot then have rotation number 1. However, this does not rule out the possibility that other types of phase locking (e.g., 2:1) might occur.

Remark. As previously noted, the shear, or radial variation in angular velocity near the limit cycle, is a property of the unforced oscillator. In our analysis of (2.1) it was assumed that β, which determines the amount of shear present in the unforced oscillator, depended on ε in such a way that $\varepsilon/\beta \to c$ as $\varepsilon \to 0$. As follows from Lemma 3.1, when $c = 0$ the angular velocity ω of solutions of the unforced oscillator is essentially constant in A_ε for ε small, since A_ε is of width $O(\varepsilon)$. On the other hand, when $c \neq 0$, ω will have an $O(1)$ variation across A_ε as $\varepsilon \to 0$. Thus, we say there is a large shear when $c \neq 0$.

The presence of large shear was accommodated in the analysis of (2.1) by making the change of variable

$$\theta = \eta + \gamma \ln r - \tan^{-1} \gamma$$

in the polar coordinate form (3.1) of (2.1). Since $r \to 1$ in A_ε and $\gamma \to c$ as $\varepsilon \to 0$, it follows that

$$\theta \to \eta - \tan^{-1} c$$

in A_ε as $\varepsilon \to 0$. Thus, we see that one effect of large shear is to appear to rotate coordinates in A_ε. In particular, the angular position of the fixed points mentioned in Theorem 5.4 will differ from that which would result from the same value of b_1 when $c = 0$ by an amount close to $\tan^{-1} c$.

The fact that $b_1 = b_0\sqrt{1+c^2}$ occurs in Theorem 5.4 instead of b_0 demonstrates that the presence of large shear can also increase the sensitivity of the oscillator to forcing. Specifically, we see that the amount of forcing required to bring about entrainment is reduced when $|c|$ is increased. Furthermore, Theorem 3.3 shows that the θ-component of the solutions of (3.2) can be approximated in the presence of large shear by a solution of (4.1) that would result from a larger forcing amplitude when $|c|$ is smaller. Since $r = 1 + O(\varepsilon)$ in A_ε, the θ-component determines the gross behavior of solutions there. Thus, not only does an increase in shear increase the sensitivity of the oscillator to forcing, but the solutions in A_ε appear to behave much as if the forcing had been increased instead of the shear.

Acknowledgment. The present paper grew out of the author's thesis. It is with pleasure that I thank my advisor Nancy Kopell, whose guidance and encouragement were essential to both this and the earlier work. I also thank G. Bard Ermentrout for helpful comments and suggestions.

REFERENCES

[1] N. KOPELL AND L. N. HOWARD, *Plane wave solutions to reaction-diffusion equations*, Stud. Appl. Math., 52 (1973), pp. 291–328.

[2] F. C. HOPPENSTEADT AND J. P. KEENER, *Phase locking of biological clocks*, J. Math. Biol., 15 (1982), pp. 339–349.

[3] W. L. KATH, *Resonance in periodically perturbed Hopf bifurcation*, Stud. Appl. Math., 65 (1981), pp. 95–112.

[4] N. G. LLOYD, *On the non-autonomous van der Pol equation with large parameter*, Proc. Cambridge Philos. Soc., 72 (1972), pp. 213–227.
[5] M. LEVI, *Qualitative analysis of the periodically forced relaxation oscillations*, Mem. Amer. Math. Soc., 32 (1981).
[6] P. HARTMAN, *Ordinary Differential Equations*, 2nd ed., Birkhauser, Boston, 1982.
[7] A. KELLEY, *The stable, center-stable, center, center-unstable, and unstable manifolds*, in Transversal Mappings and Flows, Benjamin, New York, 1967, Appendix C.
[8] V. A. PLISS, *Non-Local Problems of the Theory of Oscillations*, Academic Press, New York, 1966.
[9] G. B. ERMENTROUT, private communication.
[10] G. B. ERMENTROUT AND N. KOPELL, *Parabolic bursting in an excitable system coupled with a slow oscillation*, SIAM J. Appl. Math., 46 (1986), pp. 233–253.
[11] B. VAN DER POL AND M. J. O. STRUTT, *On the stability of solutions of Mathieu's equation*, Philosophical Magazine, 7 (1928), pp. 18–38.
[12] M. I. WEINSTEIN AND J. B. KELLER, *Hill's equation with a large potential*, SIAM J. Appl. Math., 45 (1985), pp. 200–214.
[13] L. N. HOWARD AND N. KOPELL, *Slowly varying waves and shock structures in reaction-diffusion equations*, Stud. Appl. Math., 56 (1977), pp. 95–145.
[14] J. MOSER, *Stable and Random Motions in Dynamical Systems*, Princeton University Press, Princeton, NJ, 1973.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

012

SECOND ORDER NONLINEAR FORCED OSCILLATIONS*

JAMES S. W. WONG†

Abstract. We study the oscillatory behaviour of solutions of second order nonlinear differential equation $x''+a(t)f(x)=g(t)$ on the half line $[0,\infty)$. Conditions on $a(t)$, $f(x)$ and $g(t)$ are given so that all solutions are oscillatory. These results represent further improvements on those given by Kartsatos, Kusano and Onose and Foster.

Key words. second order, nonlinear, differential equations, oscillation

AMS(MOS) subject classifications. primary 34C10, 34C15

1. Introduction. We are concerned here with second order nonlinear differential equation on the half line $[0,\infty)$

$$x''+a(t)f(x)=g(t), \qquad t\in[0,\infty), \tag{1}$$

where $a(t)$, $g(t)$ are real-valued piecewise continuous functions on $[0,\infty)$ and $f(x)$ is a continuous and nondecreasing function of $x\in(-\infty,\infty)$. We shall assume that functions $a(t)$, $g(t)$ and $f(x)$ are sufficiently smooth so that equation (1) always has solutions that are continuable throughout $[0,\infty)$. Such a solution is said to be *oscillatory* if it has arbitrarily large zeros, i.e., for any $T>0$ there exists a $t\geqq T$ such that $x(t)=0$. Otherwise, the solution is said to be *nonoscillatory*, i.e., it is eventually positive or negative. Equation (1) is said to be *oscillatory* if all continuable solutions are oscillatory. Our interest is to find conditions on $a(t)$, $f(x)$ and $g(t)$ so that (1) is oscillatory.

In an earlier paper [29], we posed the problem of whether (1) remains oscillatory subject to a periodic forcing term $g(t)$, i.e., $g(t+P)=g(t)$ for all t and some positive constant P, provided that its unforced equation

$$u''+a(t)f(u)=0, \qquad t\in[0,\infty), \tag{2}$$

is oscillatory, where $f(x)$ satisfies, in addition,

$$xf(x)>0, \qquad x\neq 0. \tag{3}$$

Subsequently, Kartsatos [12] and Teufel [28] proved results showing that certain well-known oscillation criteria for the unforced equation (2) remain valid for (1) when $g(t)$ is periodic.

Kartsatos' technique introduced in [11], [12] is to assume the existence of a function $h(t)$ such that $h''(t)=g(t)$ and to reduce (1) to a second order homogeneous equation like (2). His technique has since been extended to higher order functional differential equations by Kartsatos and Manougian [14], Kartsatos and Onose [15], Kartsatos and Toro [16], Kusano and Onose [19], Onose [21], Staikos and Sficas [26], [27], and Foster [7], [8]. On the other hand, there are also a number of papers concerned with the more special linear equation, and the oscillatory nature of its solutions (see, e.g., Keener [17], Skidmore and Leighton [24], Skidmore and Bowers [25], Rankin [22], and Howard [9]). Other related results on forced oscillation for (1) may be found in Komkov [18] and Rankin [23].

* Received by the editors February 11, 1987; accepted for publication May 20, 1987.

† China Dyeing Works, Ltd., 819 Swire House, Hong Kong, and Department of Mathematics, University of Hong Kong, Hong Kong.

The basic assumptions in Kartsatos' results [11], [12] are that $h(t)$ is either small for large values of t or it is periodic in t. It is useful to note that Atkinson [2] showed that if $g(t)$ is the second derivative of a periodic function, then a periodic second primitive $h(t)$ exists such that $h''(t)=g(t)$. Howard's results for the linear equation apply however to the more general oscillatory functions such as $g(t)=t^{\delta}\sin t$, δ being any real number.

More specifically, we consider the following specific Emden-Fowler equation:

$$x''+t^{\alpha}|x|^{\gamma}\operatorname{sgn} x=t^{\delta}\sin t, \qquad \gamma>0, \tag{4}$$

on $[0,\infty)$ where δ is any real number. Using Kartsatos' theorems [11], [12], we can conclude that (4) is oscillatory if $\alpha\geqq -1$ and $\delta\leqq 0$. Howard's theorem [9], when applied to a special case of (4), i.e.,

$$x''+x=t^{\delta}\sin t, \quad \delta \text{ real},$$

does yield oscillation for all values of δ. The results given below represent a further improvement in this direction and will show among other things that (4) is oscillatory for all values of δ provided that $\alpha+\gamma\delta>-1$. In the special cases when $\gamma>1$ and $0<\gamma<1$, sharper conditions are also available. Applications of our results to specific examples may be found in § 4.

2. Throughout this paper, we assume that $f(x)$ is continuous and nondecreasing in x satisfying condition (3), and that $a(t)$ is nonnegative but not eventually zero on $[0,\infty)$. Furthermore, we assume the following hypothesis on the forcing term:

(H_1) There exists an $h(t)\in C^2[0,\infty)$ such that $h''(t)=g(t)$ and that $h(t)$ is oscillatory, i.e., it has unbounded zeros.

Let $x(t)=y(t)+h(t)$, then (1) can be rewritten as a homogeneous equation

$$y''+a(t)f(y+h)=0. \tag{5}$$

To prove (1) is oscillatory, it is sufficient to assume the existence of an eventually positive solution $x(t)$ and deduce a contradiction by applying the various hypothesis to (5). Suppose that $x(t)>0$ on $[t_0,\infty)$. Since $a(t)\geqq 0$, from (5) we note that $y''(t)\leqq 0$ on $[t_0,\infty)$. In our first step, we show that $y'(t)\geqq 0$ on $[t_1,\infty)$ for some $t_1\geqq t_0$. If not, say $y'(t_2)<0$ for some $t_2\geqq t_0$. Since $y''(t)\leqq 0$, $y'(t)\leqq y'(t_2)<0$ for all $t\geqq t_2$; hence $y(t)\to-\infty$ as $t\to\infty$, but this together with $h(t)$ being oscillatory contradicts the assumption that $x(t)>0$. In fact, given that $y''(t)\leqq 0$, $y'(t)\geqq 0$, we must have $y'(t)$ eventually positive, i.e., $y'(t)>0$ for $t\geqq t_3\geqq t_0$. Suppose that $y'(t_3)=0$ for some $t_3\geqq t_0$. By $y''\leqq 0$, $y'\geqq 0$, this means that $y'(t)\equiv 0$ for all $t\geqq t_3$, and $y''(t)\equiv 0$. Returning to (5), this would imply $a(t)\equiv 0$ on $[t_3,\infty)$, contradicting the assumption stated at the beginning of this section.

Next, we show that $y(t)$ is eventually positive. Since $x(t)>0$ and $h(t)$ is oscillatory, so $y(t)=x(t)+h(t)$ certainly cannnot be eventually negative, nor can it be identically zero. On the other hand, if $y'(t)\geqq 0$ on $[t_1,\infty)$, then $y(t)$ certainly cannot be oscillatory. Hence, we must have that $y(t)>0$, $t\in[t_4,\infty)$, for $t_4\geqq t_1$. Thus, for simplicity, we conclude that $y(t)>0$, $y'(t)>0$ and $y''(t)\leqq 0$ eventually hold on $[0,\infty)$. We shall repeatedly avail to this conclusion in proving the various results in this paper.

THEOREM 1. *Assume that* (H_1) *holds and that* $h(t)$ *satisfies, in addition,*

$$\liminf_{t\to\infty} t^{-1}h(t)=-\infty \quad \textit{and} \quad \limsup_{t\to\infty} t^{-1}h(t)=+\infty. \tag{H_2}$$

Then (1) *is oscillatory.*

Proof. Under the given hypothesis, we have that $y(t)>0$, $y'(t)>0$ and $y''(t)\leqq 0$, $t\in[t_0,\infty)$ for some $t_0\geqq 0$. This implies that there exists a constant $M>0$ such that $0<y(t)\leqq Mt$ for large t, or

$$\limsup_{t\to\infty} t^{-1}y(t)\leqq M. \tag{6}$$

On the other hand, we have that $y(t)+h(t)=x(t)>0$ for large t, or $y(t)>-h(t)$. Dividing t and taking lim sup on both sides of $y>-h$, we immediately obtain a contradiction by invoking (H_2) to the inequality (6). Here we note that

$$\limsup_{t\to\infty} -t^{-1}h(t)=-\liminf_{t\to\infty} t^{-1}h(t)=+\infty. \tag{7}$$

The other part of hypothesis (H_2) is required when we assume the nonoscillatory solution $x(t)$ to be eventually negative and use a similar equation to (7) in that case.

THEOREM 2. *Assume that* (H_1) *holds and, in addition, that* $h(t)$ *satisfies*

$$\int_0^\infty a(t)f(h_+(t))\,dt=\int_0^\infty a(t)f(h_-(t))\,dt=+\infty \tag{H_3}$$

where $h_+(t)=\max\{h(t),0\}$ *and* $h_-(t)=\min\{h(t),0\}$. *Then* (1) *is oscillatory.*

Proof. As before, we may assume that $y(t)>0$, $y'(t)>0$ and $y''(t)\leqq 0$ on $[t_0,\infty)$. Integrating (5), we obtain

$$y'(t)-y'(t_0)+\int_{t_0}^t a(s)f(y(s)+h(s))\,ds=0. \tag{8}$$

Since $y''\leqq 0$, $\lim_{t\to\infty} y'(t)$ exists and is finite; hence the integral in (8) converges as $t\to\infty$.

We note that for all $t\geqq t_0$, $y(t)+h(t)>h_+(t)$. To see this, we write $y+h=y+h_+-h_-$ and observe that

(i) for $h_+=0$, $y+h=y-h_-=x>0=h_+$, and

(ii) for $h_-=0$, $y+h=y+h_+>h_+$ (since $y>0$).

Since f is nondecreasing, we have that $f(y+h)\geqq f(h_+)$. With $a(t)\geqq 0$, we can estimate as follows:

$$\int_{t_0}^t a(s)f(h_+(s))\,ds\leqq\int_{t_0}^t a(s)f(y+h)\,ds<\infty. \tag{9}$$

By applying (H_3) to (9), we obtain the desired contradiction.

THEOREM 3. *Assume that* (H_1) *holds and, in addition, that* $a(t)\geqq 0$ *with* $\int_0^\infty a(t)\,dt=\infty$, *and* $h(t)$ *satisfies*:

$$|h(t)|\leqq M, \text{ and } \lim_{t\to\infty} h(t) \text{ does not exist.} \tag{H_4}$$

Then the derivative of every solution of (1) *is oscillatory. Furthermore, all unbounded solutions are oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (1), say $x(t)>0$ on $[t_0,\infty)$. We first show that if $x(t)$ is unbounded, then it must be oscillatory. Following the same argument as before, we may assume that $y(t)>0$, $y'(t)>0$, and $y''(t)\leqq 0$ on $[t_0,\infty)$. By (H_4), $|h(t)|\leqq M$. Suppose that $x(t)$ is unbounded, then $y(t)$ must also be unbounded. Otherwise $x(t)=y(t)-h(t)$ becomes bounded. Let $t_1\geqq t_0$ be such that $y(t_1)=2M$; then $y(t)\geqq 2M$ for all $t\geqq t_1$. Now, $y(t)+h(t)\geqq 2M-h_-(t)\geqq M$, hence $f(y+h)\geqq f(M)$ which, upon substituting in (8) and using $\int_0^\infty a=\infty$, we obtain the desired contradiction.

We now suppose that $x(t)$ is bounded and show that $x'(t)$ must be oscillatory. Since $h(t)$ is bounded, so is $y(t)$. Note that $y'' \leqq 0$ implies that $y'(t) \to 0$ and $y(t) \to c$ as $t \to \infty$, where c is some positive constant. If $x'(t)$ is eventually of one sign, it cannot be $x'(t) < 0$ because $y'(t) + h'(t) < 0$ would contradict $y' > 0$ when we set t equal to any zero of $h'(t)$. On the other hand, if $x'(t) > 0$, then $\lim_{t\to\infty} x(t) = b$ for some positive constant. Hence, $h(t) = x(t) - y(t)$ tends to $b - c$ as $t \to \infty$. This clearly contradicts (H_4). The proof is now complete.

THEOREM 4. *Assume that* (H_1) *holds and* $a(t) \geqq 0$ *with* $\int_0^\infty a = \infty$. *Suppose, in addition, that* $h(t)$ *satisfies*:

(H_5) *There exist sequence* $\{s_n\}, \{\bar{s}_n\}$ *such that* $\lim_{n\to\infty} s_n = \lim_{n\to\infty} \bar{s}_n = \infty$ *as* $n \to \infty$, *and* $h(s_n) = \inf\{h(t): t \geqq s_n\}$ $h(\bar{s}_n) = \sup\{h(t): t \geqq \bar{s}_n\}$.

Then (1) *is oscillatory.*

Proof. Once again we begin with $y(t) > 0$ and $y'(t) > 0$ on $[t_0, \infty)$. Note that there exists n_0 such that $s_{n_0} \geqq t_0$ and for $t \geqq s_{n_0} \geqq t_0$,

$$y(t) + h(t) \geqq y(s_{n_0}) + h(s_{n_0}) = x(s_{n_0}) > 0. \tag{10}$$

Substituting (10) into (8) and using the fact that f is nondecreasing and $\int_0^\infty a = \infty$, we find that $y'(t) \to -\infty$ as $t \to \infty$ which clearly contradicts $y'(t) > 0$ on $[t_0, \infty)$.

COROLLARY 1. *Suppose that* (H_1) *holds,* $a(t) \geqq 0$ *with* $\int_0^\infty a = \infty$, *and* $h(t)$ *satisfies the condition that* $\lim_{t\to\infty} h(t) = 0$ *or* $h(t)$ *is periodic in* t. *Then* (1) *is oscillatory.*

If $h(t) \to 0$ as $t \to \infty$ or $h(t)$ is periodic, then it is easy to see that (H_5) is satisfied; hence Corollary 1 follows from Theorem 4.

3. In this section, we shall consider the so-called superlinear and sublinear equations. In the case of the Emden-Fowler equation (4), this corresponds to $\gamma > 1$ and $0 < \gamma < 1$, respectively. For the more general equations (1) and (2), we said it is *superlinear* if

$$\int_\varepsilon^\infty \frac{dx}{f(x)} < \infty, \quad \int_{-\varepsilon}^{-\infty} \frac{dx}{f(x)} < \infty \quad \text{for any } \varepsilon > 0, \tag{11}$$

and it is *sublinear* if

$$\int_0^\varepsilon \frac{dx}{f(x)} < \infty, \quad \int_0^{-\varepsilon} \frac{dx}{f(x)} < \infty \quad \text{for any } \varepsilon > 0. \tag{12}$$

For the unforced equation (2), in the superlinear case Macki and Wong [20] have extended a well-known result of Atkinson [1] by proving the condition that $a(t) \geqq 0$ and

$$\lim_{T\to\infty} \int_0^T t a(t)\, dt = \infty \tag{13}$$

is necessary and sufficient for the oscillation of (2). Kamenev [10] shows that condition (13) alone suffices for oscillation, if we drop the assumption that $a(t) \geqq 0$. We now show that Theorem 4 can be improved in the superlinear case by relaxing the assumption that $\int_0^\infty a = \infty$ to (13).

THEOREM 5. *Assume that* $h(t)$ *satisfies* (H_1) *and* (H_5), *and* $a(t) \geqq 0$ *satisfying* (13). *Suppose, in addition, that* $f(x)$ *satisfies* (11). *Then* (1) *is oscillatory.*

Proof. We begin with $x(t) > 0$, $y(t) > 0$, $y'(t) > 0$ and $y''(t) \leqq 0$ on $[t_0, \infty)$. By (H_5), there exists n_0 such that for $t \geqq s_{n_0} \geqq t_0$ we have that

$$y(t) + h(t) \geqq y(t) + h(s_{n_0}) = z(t), \tag{14}$$

when $z(t)$ is defined by (14). Since $h(s_{n_0})$ is constant, from (5) we derive

$$z''+a(t)f(z)\leqq y''+a(t)f(y+h)=0 \tag{15}$$

because $z'=y'$ and $z''=y''$. Moreover, for $t\geqq s_{n_0}$,

$$z(t)=y(t)+h(s_{n_0})\geqq y(s_{n_0})+h(s_{n_0})=x(s_{n_0})>0. \tag{16}$$

Thus, we can now reduce the proof to the second order differential inequality (15) and apply the original method of proof for (2) as given in [20] to arrive at a desired contradiction. For the sake of completeness, we adopt a somewhat simpler argument developed in later papers (see, e.g., [19], [30]). By (16), we can divide (15) by $(f(z))^{-1}$ and then multiply through by t to arrive at

$$\frac{tz''}{f(z)}+ta(t)=0. \tag{17}$$

Now, integrate (17) from some $t_1\geqq s_{n_0}$ to t and obtain

$$\int_{t_1}^{t}\frac{sz''}{f(z)}ds+\int_{t_1}^{t}sa(s)\,ds=0. \tag{18}$$

Carrying out the first integral in (18), we have that

$$\int_{t_1}^{t}\frac{sz''}{f(z)}=\frac{tz'(t)}{f(z(t))}-\frac{t_1z'(t_1)}{f(z(t_1))}+\int_{t_1}^{t}\frac{sz'^2f'(z)}{f^2(z)}-\int_{t_1}^{t}\frac{z'}{f(z)}. \tag{19}$$

Since the second integral on (19) above is nonnegative, and we also have that $z'(t)>0$, we can estimate (18) by dropping these two terms as follows:

$$\int_{t_1}^{t}sa(s)\,ds\leqq\frac{t_1z'(t_1)}{f(z(t_1))}+\int_{z(t_1)}^{z(t)}\frac{d\xi}{f(\xi)}. \tag{20}$$

The last integral is finite by (11) and the fact that $z'>0$, so (13) produces the desired contradiction in (20) upon letting t tend to ∞. We remark that in (19) we have loosely used the term $f'(z)$, although differentiability of f has not been assumed. However, since f is nondecreasing we can make the integration by parts into a rigorous proof by approximations. Another approach is to use Lebesgue-Stieltjes integrals.

Next we turn our attention to the sublinear equation and refer to the corresponding results of Belohorec. In the special case of $f(x)=|x|^{\gamma}\operatorname{sgn}x$, $0<\gamma<1$, Belohorec [4] proved the analogue of Atkinson's theorem: the condition that $a(t)\geqq 0$ and

$$\lim_{T\to\infty}\int_0^T t^{\gamma}a(t)\,dt=\infty, \tag{21}$$

is necessary and sufficient for the oscillation of (2). Subsequently, Belohorec [5] also showed that condition (21) also remains sufficient for oscillation without $a(t)\geqq 0$. However, extension to (2) in its more general form, i.e., f satisfying (12), remains elusive. An attempt was made by Coles [6]; however, the additional assumption required on f, which states that there exists a positive constant c such that for all x

$$f'(x)\int_0^x\frac{d\xi}{f(\xi)}\geqq c>0, \tag{22}$$

seems somewhat artificial. On the other hand, Coles indicates at the end of his paper [6] that the natural extension of divergence condition (21) should perhaps be

$$\lim_{T\to\infty}\int_0^T f(t)a(t)\,dt=\infty. \tag{23}$$

In our next result we give an analogue of Theorem 5 in the sublinear case, except that ours is subject to additional conditions on f that f is even, $f(-x)=-f(x)$ and also

$$f(uv)\geqq f(u)f(v) \quad \text{for all } u>0, \text{ and } v \text{ large.} \tag{24}$$

Condition (24) may be referred to as supermultiplicativeness. Aside from rather restricted nature of this assumption, which is clearly satisfied by $f(x)=|x|^{\gamma}\sin x$, $0<\gamma<1$, our result does extend Belohorec's original result [4] when applied to (1) with $g(t)\equiv 0$.

THEOREM 6. *Assume that $h(t)$ satisfies (H_1) and (H_5), and $a(t)\geqq 0$ satisfying (23). Suppose, in addition, that $f(x)$ satisfies (12) and (24). Then (1) is oscillatory.*

Proof. We follow the argument in Theorem 5 up to the second order differential inequality

$$z''(t)+a(t)f(z(t))\leqq 0, \qquad t\geqq t_0, \tag{15}$$

with $z(t)>0$, $z'(t)>0$ and $z''(t)\leqq 0$ on $[t_0,\infty)$. Since $z''\leqq 0$, we have for some $0<\lambda<1$,

$$z(t)-z(t_0)=\int_{t_0}^{t} z'(s)\,ds\geqq z'(t)(t-t_0)\geqq \lambda z'(t)t. \tag{25}$$

Using (25) in (15) above, we obtain

$$z''(t)+a(t)f(\lambda z'(t)t)\leqq 0, \qquad t\geqq t_0$$

which by the supermultiplicative property (24) of f may be further reduced to

$$z''(t)+a(t)f(t)f(\lambda z'(t))\leqq 0. \tag{26}$$

Dividing the above through by $f(\lambda z'(t))$ and integrating, we first note that

$$\int_{t_0}^{t}\frac{z''(s)\,ds}{f(\lambda z'(s))}=\frac{1}{\lambda}\int_{\lambda z'(t_0)}^{\lambda z'(t)}\frac{d\xi}{f(\xi)},$$

from which we can estimate (26) as follows:

$$\int_{t_0}^{t} a(s)f(s)\,ds\leqq\frac{1}{\lambda}\int_{\lambda z'(t)}^{\lambda z'(t_0)}\frac{d\xi}{f(\xi)}<\infty. \tag{27}$$

The last integral in (27) is finite in view of (12) and the fact that z' is nonincreasing. The desired contradiction thus follows by applying (23) to (27).

4. In this section, we first apply the results given in the previous two sections to the specific example:

$$x''+t^{\alpha}|x|^{\gamma}\operatorname{sgn} x=t^{\delta}\sin t \qquad \gamma>0. \tag{4}$$

In (4), $g(t)=t^{\delta}\sin t$. A convenient second primitive $h(t)$ can be chosen with the following asymptotic behaviour:

$$h(t)=t^{\delta}\sin t+O(t^{\delta-1}), \qquad t\to\infty. \tag{28}$$

Applying Theorem 1 to (4), we deduce from (28) that (4) is oscillatory for all $\delta>1$ and for all values of α. Next, we apply Theorem 2 to (4), and again by using (29) in (H_4), we conclude that (4) is oscillatory if $\alpha+\gamma\delta>-1$. So if $\alpha\geqq -1$, $\delta>0$, then hypothesis (H_3) is satisfied and we have oscillation.

Last, we consider Theorem 4 for this specific example and note that for $\delta\leqq 0$, $h(t)=t^{\delta}\sin t$ satisfies (H_5), so we have oscillation of (4) if $\alpha\geqq -1$ and $\delta\leqq 0$. This conclusion is complementary to that determined by Theorem 2 and can be obtained

also by applying the original results of Kartsatos [11], [12]. For the superlinear case, $\gamma > 1$, we can use Theorem 5 to improve the condition on α and conclude oscillation for $\alpha \geqq -2$ and $\delta \leqq 0$. Likewise in the sublinear case, with $0 < \gamma < 1$, we use Theorem 6 to deduce oscillation (4) for $\alpha \geqq -(1+\gamma)$ and $\delta \leqq 0$.

Next, consider the following example:

$$x'' + t^{\alpha} x\, e^{|x|^2} = e^{-t} \sin t, \tag{29}$$

which arises from a certain radial solution of the Klein-Gordon equation in physics (see Atkinson and Peletier [3] and also Wong [31]). This example is also discussed in detail by Rankin [23], but the conditions required on $g(t)$ seem almost artificial. In any case, Rankin's result, when applied to (29), yields oscillation for $\alpha \geqq -1$. Here, $h(t) = \frac{1}{2}e^{-t} \cos t$, which tends to zero as $t \to \infty$. An application of Theorem 5 will result in oscillation of (29) for all $\alpha \geqq -2$ in this superlinear case.

As a third example, we consider the following sublinear equation, with $0 < \gamma < 1$,

$$x'' + t^{\alpha}(1+\sin t)|x|^{\gamma} \operatorname{sgn} x = \sin t\left(1 + \frac{1}{t} - \frac{2}{t^3}\right) + \frac{2\cos t}{t^2}. \tag{30}$$

First note that $h(t) = -1 + (1/t)\sin t$, which satisfies (H_5) (here $h(t)$ is neither periodic nor does it tend to zero), and then apply Theorem 6 to conclude oscillation of (30) for $\alpha \geqq -(1+\gamma)$.

For concluding remarks, we first note that Kartsatos [11], [12] and others stated their results for the more general nth order equations. Kusano and Onose [19] and others extended further to include delay differential equations. Some of our results admit ready extensions to higher order equations and others may not. We hope to return to such a discussion at a later data. Second, our results are based mainly on known techniques for nonlinear oscillations with nonnegative "potential" $a(t)$, and little is known concerning what occurs when $a(t)$ is allowed to change signs even in the linear case. Howard [9] made an extensive investigation using techniques for the linear equation without explicitly requiring $a(t)$ to be nonnegative. Unfortunately, his results do not apply to the many interesting examples when $a(t)$ does change sign. In particular, we are unable to determine whether the following simple linear is oscillatory:

$$x'' + (1 + 2\cos t)x = \sin t.$$

The solution to such a question will represent a major step forward in the study of forced oscillations. Third, we remark that our results depend heavily on the assumption that $f(x)$ is nondecreasing. It will be useful to prove results which do not require $f(x)$ to be monotone.

Returning to Theorem 6, we claimed that in the special case when $g(t) \equiv 0$ (which obviously satisfies (H_5)) it is an extension of Belohorec's result [4] originally proved for $f(x) = |x|^{\gamma} \operatorname{sgn} x$, $0 < \gamma < 1$. However, the supermultiplicative assumption (24) is rather restrictive and seems no better than Coles' condition (22). We believe that such an extension remains valid without the additional assumption (24), and in any case, one should expect it to be replaced by some less restrictive conditions. Finally, we began this paper by referring to a problem posed nearly twenty years ago, on the oscillation of (1) subject to a periodic forcing term. The original question stated in [29] was in fact (1) with $g(t) = \sin t$, i.e., a periodic function with zero mean value. The same question can now be extended by requiring $g(t)$ to be only an almost periodic function with mean value zero.

In closing, we refer the reader to the survey article by Kartsatos [13] and two others by the author [30], [32], where additional references and other open problems may be found.

REFERENCES

[1] F. V. ATKINSON, *On second order nonlinear oscillation*, Pacific J. Math., 5 (1955), pp. 643-647.

[2] ———, *On second order differential inequalities*, Proc. Roy. Soc. Edinburgh sect A, 72 (1972/3), pp. 109-127.

[3] F. V. ATKINSON AND L. A. PELETIER, *Ground states of $-\Delta u = f(u)$ and the Emden-Fowler Equation*, Arch. Rational Mech. Anal., 93 (1986), pp. 103-127.

[4] S. BELOHOREC, *Oscillatory solutions of certain nonlinear differential equation of second order*, Mat. Fyz. Casopis, Sloven. Akad. Vied., 11 (1961), pp. 250-255. (In Slovak.)

[5] ———, *Two remarks on the properties of solutions of a nonlinear differential equations*, Acta Fac. Rerum. Natur. Univ. Comen. Mathematica, XXII (1969), pp. 19-26.

[6] W. J. COLES, *A nonlinear oscillation theorem*, International Conference on Differential Equations, H. A. Antosiewic, ed., Academic Press, New York, 1975, pp. 193-202.

[7] K. FOSTER, *Criteria for oscillation and growth of nonoscillatory solutions of forced differential equations of even order*, J. Differential Equations, 20 (1976), pp. 115-132.

[8] ———, *Oscillations of forced sublinear differential equations of even order*, J. Math. Anal. Appl., 55 (1976), pp. 636-643.

[9] H. C. HOWARD, *Oscillation and nonoscillation criteria for nonhomogeneous differential equations*, Ann. Mat. Pura et Appl. (1977), pp. 163-180.

[10] L. V. KAMENEV, *Oscillation of solutions of second order nonlinear equations*, Trans. Moscow Electrical Machine Design Institute, 5 (1969/70), pp. 125-136.

[11] A. G. KARTSATOS, *On the maintenance of oscillations of nth order equations under the effect of a small forcing term*, J. Differential Equations, 10 (1971), pp. 355-363.

[12] ———, *Maintenance of oscillations under the effect of a periodic forcing term*, Proc. Amer. Math. Soc., 33 (1972), pp. 377-383.

[13] ———, *Recent results on oscillations of solutions of forced and perturbed nonlinear differential equations of even order*, in Stability of Dynamical Systems, CMBS-NSF Conference-Mississipi State University 1975, Lecture Notes in Pure and Applied Math Z8, Marcel Dekker, New York, 1977, pp. 17-72.

[14] A. G. KARTSATOS AND M. N. MANOUGIAN, *Perturbations causing oscillation of functional differential equations*, Proc. Amer. Math. Soc., 43 (1974), pp. 111-117.

[15] A. G. KARTSATOS AND H. ONOSE, *A comparison theorem for functional differential equations*, Bull. Austral. Math. Soc., 14 (1976), pp. 343-347.

[16] A. G. KARTSATOS AND J. TORO, *Oscillation and asymptotic behavior of forced nonlinear equations*, this Journal, 10 (1979), pp. 86-95.

[17] M. S. KEENER, *Solutions of certain linear nonhomogeneous 2nd order differential equations*, Applicable Anal., 1 (1971), pp. 57-63.

[18] V. KOMKOV, *On boundedness and oscillation of the differential equation $x''+A(t)g(x)=f(t)$ in R^n*, SIAM J. Appl. Math., 22 (1972), pp. 561-568.

[19] T. KUSANO AND H. ONOSE, *Oscillations of functional differential equations with retarded argument*, J. Differential Equations, 15 (1974), pp. 269-277.

[20] J. W. MACKI AND J. S. W. WONG, *Oscillation of solutions to second-order nonlinear differential equations*, Pacific J. Math., 24 (1968), pp. 111-117.

[21] H. ONOSE, *A comparison theorem and the forced oscillation*, Bull. Austral. Math. Soc., 13 (1975), pp. 13-19.

[22] S. M. RANKIN, *Oscillation theorems for second order nonhomogeneous linear differential equations*, J. Math. Anal. Appl., 53 (1976), pp. 550-553.

[23] ———, *Oscillation of a forced second order nonlinear differential equation*, Proc. Amer. Math. Soc., 59 (1976), pp. 279-282.

[24] A. SKIDMORE AND W. LEIGHTON, *On the differential equation $y''+p(x)y=f(x)$*, J. Math. Anal. Appl., 43 (1973), pp. 46-55.

[25] A. SKIDMORE AND J. J. BOWERS, *Oscillatory behavior of solutions of $y''+p(x)y=f(x)$*, J. Math. Anal. Appl., 49 (1975), pp. 317-323.

[26] V. A. STAIKOS AND Y. G. SFICAS, *Oscillations for forced second order nonlinear differential equations*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 55 (1973), pp. 25-30.

[27] V. A. STAIKOS AND Y. G. SFICAS, *Forced oscillations for differential equations of arbitrary order*, J. Differential Equations, 17 (1975), pp. 1-11.
[28] H. TEUFEL, JR., *Forced second order nonlinear oscillations*, J. Math. Anal. Appl., 40 (1972), pp. 148-152.
[29] J. S. W. WONG, *On second order nonlinear oscillation*, Funkcial Ekvac., 11 (1969), pp. 207-234.
[30] ———, *Oscillation theorems for second order nonlinear differential equations*, Bull. Inst. Math. Acad. Sinica, 3 (1975), pp. 283-309.
[31] ———, *On the generalized Emden-Fowler equation*, SIAM Rev., 17 (1975), pp. 339-360.
[32] ———, *Classifications of second order nonlinear oscillations*, to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

013

RECURSIVELY GENERATED POLYNOMIALS AND GERONIMUS' VERSION OF ORTHOGONALITY ON A CONTOUR*

JOHN W. JAYNE†

Abstract. Let $\{p_n\}$ be a sequence of polynomials generated by the three-term recurrence $p_0=1$, $p_1=z+b_0$, $p_{n+1}=(z+b_n)p_n-c_np_{n-1}$, $n\geqq 1$. Using only the hypothesis that $\{b_n\}$ and $\{c_n\}$ are bounded complex sequences (with $c_n\neq 0$), we show by constructing a weight function as a convergent Laurent series—with coefficients given explicitly in terms of b_n and c_n—that the p_n are orthogonal on a contour in the sense that Geronimus describes (and that the Bessel polynomials exemplify). We thus obtain an elementary approach to this kind of orthogonality and provide a construction of the weight function which is an alternative to the continued fraction representation discussed by Askey and Ismail.

Key words. orthogonal polynomials, three-term recurrence relation, orthogonality on a contour

AMS(MOS) subject classifications. 33A65, 30E05

1. Introduction. In order to establish a frame of reference for this paper we review several points concerning orthogonal polynomials. Let Ψ be a real-valued, nondecreasing function on $(-\infty,\infty)$, all of whose moments $\int_{-\infty}^{\infty}x^n\,d\Psi$ exist. A sequence of polynomials $\{p_n\}$—assumed without loss of generality to be monic—is orthogonal (on the real line) with respect to Ψ if

$$\int_{-\infty}^{\infty}p_n(x)p_m(x)\,d\Psi(x)=\begin{cases}0, & m\neq n,\\ h_n>0, & m=n\geqq 0.\end{cases}\tag{1}$$

It has long been known that a sequence orthogonal in this sense satisfies a three-term recurrence relation

$$\begin{aligned}&p_0=1,\\ &p_1=x+b_0,\\ &p_{n+1}=(x+b_n)p_n-c_np_{n-1},\qquad n\geqq 1,\end{aligned}\tag{2}$$

where b_n is real and $c_n>0$. An early version of a converse to this statement is usually attributed to Favard. In its present form, which is due to the development of certain representation theorems for moment functionals, it runs as follows: if $\{p_n\}$ is generated by (2) where $\{b_n\}$ and $\{c_n\}$ are arbitrary *complex* sequences with $c_n\neq 0$, then there exists a function Ψ of bounded variation on $(-\infty,\infty)$ such that (1) holds, with $h_n\neq 0$. Furthermore Ψ can be chosen to be real-valued if and only if $\{b_n\}$ and $\{c_n\}$ are real sequences; Ψ is nondecreasing with infinite spectrum if and only if b_n is real, $c_n>0$. For a complete discussion see Chihara [2].

The concept of orthogonality in (1) is of course not the only one possible. Another for which (2) also holds (with complex coefficients and $c_n\neq 0$) is that of orthogonality on a simple closed curve C, as discussed by Geronimus [3]: instead of the distribution Ψ there is assumed to be a weight function w having one or more singularities inside C and such that

$$(1/2\pi i)\int_C p_n(z)p_m(z)w(z)\,dz=\begin{cases}0, & m\neq n,\\ h_n\neq 0, & m=n\geqq 0.\end{cases}\tag{3}$$

Perhaps the best known example is the sequence of Bessel polynomials, for which w has an essential singularity at the origin [4]. However, the Bessel polynomials are not

* Received by the editors March 7, 1986; accepted for publication May 18, 1987.
† Department of Mathematics, University of Tennessee, Chattanooga, Tennessee 37403.

unique in having this type of orthogonality: Geronimus showed that a sequence of polynomials which is orthogonal in the sense of (1) on a finite interval is also orthogonal in the sense of (3).

A problem to which a great deal of research has been devoted is that of recovering the distribution function Ψ from the recurrence (2). In [1] Askey and Ismail treat this problem by showing that under certain conditions w exists as a continued fraction obtainable from (2), and then they derive a fundamental relation between w and Ψ from which Ψ may be determined once w is found. Motivated by the example of the Bessel polynomials we show in this paper that w can be obtained directly from (2) as a convergent Laurent series, with coefficients given by explicit formulas in terms of b_n and c_n. Our only hypothesis is that $\{b_n\}$ and $\{c_n\}$ are bounded complex sequences with $c_n \neq 0$—we do not make any use of the theory of continued fractions (nor refer to orthogonality on the real line). Whether this approach would be helpful in recovering Ψ remains to be seen; it does provide an elementary proof that many families of recursively generated polynomials are orthogonal in the sense of (3), itself a fact of considerable interest.

2. Preliminaries: formal construction of the weight function. We start with the recurrence (2), in which b_n and c_n are taken to be complex, with $c_n \neq 0$. We assume for the present that (3) holds and furthermore that w has the representation

$$w(z) = \sum_{j=1}^{\infty} w_j z^{-j}, \qquad w_1 = 1 \tag{4}$$

for all z outside some circle $|z| = R$ lying interior to C. Our immediate task is to calculate the coefficients w_j in (4). One way to approach this problem is the following. Let the coefficient of z^k in p_n be denoted by $a(n, k)$ for $n \geqq 0$, $0 \leqq k \leqq n$ (note that $a(n, n) = 1$ since the polynomials are monic) and set $a(n, k) = 0$ if $k > n$. Since $p_0 = 1$ it follows from (3) that

$$(1/2\pi i) \int_C p_n(z) w(z)\, dz = \begin{cases} 0, & n \geqq 1, \\ h_0 \neq 0, & n = 0. \end{cases} \tag{5}$$

Inserting (4) and the representation

$$p_n(z) = \sum_{k=0}^{n} a(n, k) z^k$$

into (5) and integrating term by term yields

$$\begin{aligned} w_1 &= h_0 = 1 \\ w_2 &= -a(1, 0) w_1 \\ w_3 &= -a(2, 0) w_1 - a(2, 1) w_2 \\ &\vdots \\ w_{n+1} &= -\sum_{k=1}^{n} a(n, k-1) w_k, \qquad n \geqq 1. \end{aligned} \tag{6}$$

We can thus compute all the coefficients w_j from (6). Recall that the moment μ_n of the weight function w is given by

$$\mu_n = (1/2\pi i) \int_C z^n w(z)\, dz,$$

so if w has the representation (4) then

$$\mu_n = (1/2\pi i)\int_C \sum_{j=1}^{\infty} w_j z^{n-j}\,dz = w_{n+1}, \qquad n \geqq 0.$$

But we can write z^n as a linear combination of the polynomials $p_0, p_1, \cdots, p_n$ generated by (2)

$$z^n = \sum_{k=0}^{n} \gamma(k, n)p_k(z) \tag{7}$$

so that we also have

$$\begin{aligned} w_{n+1} = \mu_n &= (1/2\pi i)\int_C z^n w(z)\,dz \\ &= (1/2\pi i)\int_C \sum_{k=0}^{n} \gamma(k, n)p_k(z)w(z)\,dz \\ &= \gamma(0, n), \qquad n \geqq 0. \end{aligned} \tag{8}$$

It is this representation, rather than (6), that will be used to derive our results, and in doing so we shall find a way to calculate *all* the $\gamma(k, n)$ (cf. [5, pp. 45ff] for related calculations).

So far, by assuming the orthogonality expressed in (3) and the existence of a weight function w given by (4), we have shown that the coefficients w_j are given by (6) or (8). If we now take (8) as our starting point and formally construct the series (4) by letting $w_{n+1} = \gamma(0, n)$ for $n \geqq 0$ we of course have no a priori guarantee that it will converge (the convergence question being our main concern), but when it does the following theorem applies (cf. the proof of Favard's Theorem [2, pp. 21-22]).

THEOREM 1. *Let* $\{p_n\}$ *be generated by* (2) *and suppose the series* (4) *is constructed by setting* $w_{n+1} = \gamma(0, n)$ *for* $n \geqq 0$. *If this series converges for all* z *outside some circle* $|z| = R$ *and* C *is any simple closed curve having* $|z| = R$ *in its interior, then*

$$(1/2\pi i)\int_C p_n(z)p_m(z)w(z)\,dz = \begin{cases} 0, & m \neq n, \\ 1, & m = n = 0, \\ \prod_{j=1}^{n} c_j, & m = n \geqq 1. \end{cases} \tag{9}$$

Proof. The case $m = n = 0$ is clear, so assume $n \geqq 1$. Then it suffices to show that

$$(1/2\pi i)\int_C z^m p_n(z)w(z)\,dz = \begin{cases} 0, & 0 \leqq m < n, \\ \prod_{j=1}^{n} c_j, & m = n \geqq 1. \end{cases}$$

If $m = 0$ the result follows from the way w was constructed; if $m = 1$ then, since (2) implies

$$zp_n(z) = p_{n+1}(z) - b_n p_n(z) + c_n p_{n-1}(z),$$

we have

$$(1/2\pi i)\int_C zp_n(z)w(z)\,dz = (c_n/2\pi i)\int_C p_{n-1}(z)w(z)\,dz = \begin{cases} 0, & n > 1, \\ c_1, & n = 1. \end{cases}$$

The conclusion follows by induction.

In view of Theorem 1 it is clear that, given a sequence $\{p_n\}$ generated by (2), we can establish orthogonality in the sense of (3) *and* simultaneously exhibit the weight function by simply constructing the series (4) and proving it converges outside some circle $|z| = R$. Since we intend to establish this convergence by assuming that $\{b_n\}$ and $\{c_n\}$ are bounded complex sequences, we need a formula for w_j explicitly in terms of b_n and c_n, and we shall obtain it from (8) with the aid of the following lemma.

LEMMA 1. *Let* $\{p_n\}$ *be generated by* (2), *and consider the representation* (7)

$$z^n = \sum_{k=0}^{n} \gamma(k, n) p_k(z).$$

With the understanding that $\gamma(i, j) = 0$ *if* $i > j$ *or* $i < 0$, *we have*

$$\gamma(k, n+1) = \gamma(k-1, n) - b_k \gamma(k, n) + c_{k+1} \gamma(k+1, n), \tag{10}$$

valid for $0 \leqq k \leqq n+1$, $n \geqq 0$. *In particular,*

$$\gamma(n, n) = 1,$$

$$\gamma(0, n+1) = -b_0 \gamma(0, n) + c_1 \gamma(1, n),$$

$$\gamma(n, n+1) = \gamma(n-1, n) - b_n, \qquad n \geqq 0.$$

Proof. Because of (7) we have

$$\begin{aligned}
z^{n+1} &= \sum_{k=0}^{n} \gamma(k, n) z p_k(z) \\
&= \sum_{k=1}^{n} \gamma(k, n)[p_{k+1}(z) - b_k p_k(z) + c_k p_{k-1}(z)] + \gamma(0, n)[p_1(z) - b_0] \\
&= \sum_{j=2}^{n+1} \gamma(j-1, n) p_j(z) - \sum_{j=1}^{n} b_j \gamma(j, n) p_j(z) + \sum_{j=0}^{n-1} c_{j+1} \gamma(j+1, n) p_j(z) \\
&\quad + \gamma(0, n) p_1(z) - \gamma(0, n) b_0 p_0(z) \\
&= \gamma(n, n) p_{n+1}(z) + [\gamma(n-1, n) - b_n \gamma(n, n)] p_n(z) \\
&\quad + \sum_{j=1}^{n-1} [\gamma(j-1, n) - b_j \gamma(j, n) + c_{j+1} \gamma(j+1, n)] p_j(z) \\
&\quad + [-b_0 \gamma(0, n) + c_1 \gamma(1, n)] p_0(z).
\end{aligned}$$

But z^{n+1} also is given by

$$z^{n+1} = \sum_{j=0}^{n+1} \gamma(j, n+1) p_j(z),$$

so the conclusion follows by equating the coefficients of $p_j(z)$ in these two expressions for z^{n+1}.

Through (8) and repeated use of Lemma 1 (i.e., by solving the difference equation (10)) we can now determine the w_j in terms of b_n and c_n. Initially the results appear very involved, but after displaying the first few $\gamma(k, n)$ and introducing suitable notation, we shall be able to guess a formula for $\gamma(k, n)$, prove its validity, and then use it to prove convergence of the series $\sum w_j z^{-j}$.

We begin by making a few calculations using (10):

$$w_1=\gamma(0,0)=1, \qquad \gamma(1,1)=1,$$
$$w_2=\gamma(0,1)=-b_0, \qquad \gamma(1,2)=-b_0-b_1, \qquad \gamma(2,2)=1,$$
$$w_3=\gamma(0,2)=b_0^2+c_1, \qquad \gamma(1,3)=b_0^2+b_0b_1+b_1^2+c_1+c_2,$$
$$w_4=\gamma(0,3)=-b_0^3-2b_0c_1-b_1c_1, \qquad \gamma(2,3)=-b_0-b_1-b_2, \qquad \gamma(3,3)=1,$$

etc. Rewriting these as iterated sums, cumbersome though it may seem at first, we take a step toward recognizing the general pattern:

$$\begin{aligned}
&w_2=\gamma(0,1)=-\sum_{j_1=0}^{0} b_{j_1},\\
&w_3=\gamma(0,2)=\sum_{j_1=0}^{0}\sum_{j_2=0}^{j_1} b_{j_1}b_{j_2}+\sum_{j_1=1}^{1} c_{j_1}, \qquad \gamma(1,2)=-\sum_{j_1=0}^{1} b_{j_1},\\
&w_4=\gamma(0,3)=-\sum_{j_1=0}^{0}\sum_{j_2=0}^{j_1}\sum_{j_3=0}^{j_2} b_{j_1}b_{j_2}b_{j_3}-\sum_{j_1=1}^{1}\sum_{j_2=0}^{j_1} c_{j_1}b_{j_2}-\sum_{j_1=0}^{0}\sum_{j_2=1}^{j_1+1} b_{j_1}c_{j_2},\\
&\gamma(1,3)=\sum_{j_1=0}^{1}\sum_{j_2=0}^{j_1} b_{j_1}b_{j_2}+\sum_{j_1=1}^{2} c_{j_1}, \qquad \gamma(2,3)=-\sum_{j_1=0}^{2} b_{j_1}.
\end{aligned} \tag{11}$$

We add one more for good measure:

$$\begin{aligned}
w_5=\gamma(0,4)=&\sum_{j_1=0}^{0}\sum_{j_2=0}^{j_1}\sum_{j_3=0}^{j_2}\sum_{j_4=0}^{j_3} b_{j_1}b_{j_2}b_{j_3}b_{j_4}+\sum_{j_1=1}^{1}\sum_{j_2=0}^{j_1}\sum_{j_3=0}^{j_2} c_{j_1}b_{j_2}b_{j_3}\\
&+\sum_{j_1=0}^{0}\sum_{j_2=1}^{j_1+1}\sum_{j_3=0}^{j_2} b_{j_1}c_{j_2}b_{j_3}+\sum_{j_1=0}^{0}\sum_{j_2=0}^{j_1}\sum_{j_3=1}^{j_2+1} b_{j_1}b_{j_2}c_{j_3}+\sum_{j_1=1}^{1}\sum_{j_2=1}^{j_1+1} c_{j_1}c_{j_2}.
\end{aligned}$$

The pattern developing here can be brought out more clearly by introducing the following notation. Let $I(k,q)$ denote the set of all possible ordered k-tuples $\mathbf{p}_k=(p_1,p_2,\cdots,p_k)$ consisting of q ones and $k-q$ zeros, where $k\geqq 1$ and $0\leqq q\leqq k$ (e.g., $I(3,1)=\{(1,0,0),(0,1,0),(0,0,1)\}$). For fixed integers $N\geqq 0$, $k\geqq 1$ and fixed $\mathbf{p}_k\in I(k,q)$, let $S(N,\mathbf{p}_k)$ denote the sum

$$S(N,\mathbf{p}_k)=\sum_{j_1=p_1}^{N+p_1}\sum_{j_2=p_2}^{j_1+p_2}\sum_{j_3=p_3}^{j_2+p_3}\cdots\sum_{j_k=p_k}^{j_{k-1}+p_k}\left(\prod_{i=1}^{k} c_{j_i}^{p_i}b_{j_i}^{1-p_i}\right) \tag{12}$$

and let $S(N,\mathbf{p}_0)=1$. Furthermore, for fixed N, k and q (with $0\leqq q\leqq k$) let $\sum_{\mathbf{p}_k\in I(k,q)} S(N,\mathbf{p}_k)$ denote the sum of all the $S(N,\mathbf{p}_k)$ having $\mathbf{p}_k$ in the index set $I(k,q)$; set $\sum_{\mathbf{p}_k\in I(k,q)} S(N,\mathbf{p}_k)=0$ if $q>k$ or $N<0$. We also agree to set $\sum_{\mathbf{p}_k\in I(k,q)} S(N,\mathbf{p}_k)=1$ if $k=0$.

By way of examples we note that

$$\sum_{\mathbf{p}_2\in I(2,1)} S(N,\mathbf{p}_2)=\sum_{j_1=0}^{N}\sum_{j_2=1}^{j_1+1} b_{j_1}c_{j_2}+\sum_{j_1=1}^{N+1}\sum_{j_2=0}^{j_1} c_{j_1}b_{j_2}$$

(since $I(2,1)=\{(0,1),(1,0)\}$), and that sums over the index sets $I(k,0)$, $I(k,k)$ actually involve only one iterated sum each

$$\sum_{\mathbf{p}_k\in I(k,0)} S(N,\mathbf{p}_k)=\sum_{j_1=0}^{N}\sum_{j_2=0}^{j_1}\sum_{j_3=0}^{j_2}\cdots\sum_{j_k=0}^{j_{k-1}} b_{j_1}b_{j_2}b_{j_3}\cdots b_{j_k},$$
$$\sum_{\mathbf{p}_k\in I(k,k)} S(N,\mathbf{p}_k)=\sum_{j_1=1}^{N+1}\sum_{j_2=1}^{j_1+1}\sum_{j_3=1}^{j_2+1}\cdots\sum_{j_k=1}^{j_{k-1}+1} c_{j_1}c_{j_2}c_{j_3}\cdots c_{j_k}.$$

Every sum appearing in the representations in (11) for $\gamma(0,1)$, $\gamma(0,2)$, $\gamma(1,2)$, $\gamma(0,3)$, $\gamma(1,3)$, $\gamma(2,3)$, and $\gamma(0,4)$ has the form (12), and consequently we can now

write these as

$$w_2 = \gamma(0,1) = -\sum_{k=1}^{1} \sum_{\mathbf{p}_k \in I(k,1-k)} S(0,\mathbf{p}_k),$$

$$w_3 = \gamma(0,2) = \sum_{\mathbf{p}_2 \in I(2,0)} S(0,\mathbf{p}_2) + \sum_{\mathbf{p}_1 \in I(1,1)} S(0,\mathbf{p}_1) = \sum_{k=1}^{2} \sum_{\mathbf{p}_k \in I(k,2-k)} S(0,\mathbf{p}_k),$$

$$\gamma(1,2) = -\sum_{k=1}^{1} \sum_{\mathbf{p}_k \in I(k,1-k)} S(1,\mathbf{p}_k),$$

$$w_4 = \gamma(0,3) = -\sum_{\mathbf{p}_3 \in I(3,0)} S(0,\mathbf{p}_3) - \sum_{\mathbf{p}_2 \in I(2,1)} S(0,\mathbf{p}_2) = -\sum_{k=2}^{3} \sum_{\mathbf{p}_k \in I(k,3-k)} S(0,\mathbf{p}_k), \tag{13}$$

$$\gamma(1,3) = \sum_{\mathbf{p}_2 \in I(2,0)} S(0,\mathbf{p}_2) + \sum_{\mathbf{p}_1 \in I(1,1)} S(0,\mathbf{p}_1) = \sum_{k=1}^{2} \sum_{\mathbf{p}_k \in I(k,2-k)} S(1,\mathbf{p}_k),$$

$$\gamma(2,3) = -\sum_{k=1}^{1} \sum_{\mathbf{p}_k \in I(k,1-k)} S(2,\mathbf{p}_k),$$

$$w_5 = \gamma(0,4) = \sum_{\mathbf{p}_4 \in I(4,0)} S(0,\mathbf{p}_4) + \sum_{\mathbf{p}_3 \in I(3,1)} S(0,\mathbf{p}_3) + \sum_{\mathbf{p}_2 \in I(2,2)} S(0,\mathbf{p}_2)$$

$$= \sum_{k=2}^{4} \sum_{\mathbf{p}_k \in I(k,4-k)} S(0,\mathbf{p}_k).$$

These representations (and additional ones which fit the developing pattern here but which we omit) form the basis for conjecturing that

$$\gamma(j,M) = (-1)^{j+M} \sum_{k=M-j-[(M-j)/2]}^{M-j} \sum_{\mathbf{p}_k \in I(k,M-j-k)} S(j,\mathbf{p}_k),$$

where $[(M-j)/2]$ is the greatest integer less than or equal to $(M-j)/2$. This will be a key result (Theorem 2), but in order to prove it we need to investigate $S(N,\mathbf{p}_k)$ in more detail. As already noted above, for convenience we set

$$\sum_{\mathbf{p}_k \in I(k,q)} S(N,\mathbf{p}_k) = \begin{cases} 1 & \text{if } k=q=0, \quad N \geqq 0 \\ 0 & \text{if either } N<0, \text{ or } q>k, \text{ or } q<0. \end{cases} \tag{14}$$

The property of the sum $\sum_{\mathbf{p}_k \in I(k,q)} S(N,\mathbf{p}_k)$ which will be needed most is furnished in the following lemma.

LEMMA 2. *For all integers* $N \geqq 0$, $k \geqq 1$, $m \geqq 1$,

$$\sum_{\mathbf{p}_m \in I(m,k-m)} S(N,\mathbf{p}_m) = \sum_{\mathbf{p}_m \in I(m,k-m)} S(N-1,\mathbf{p}_m) + b_N \sum_{\mathbf{p}_{m-1} \in I(m-1,k-m)} S(N,\mathbf{p}_{m-1}) + c_{N+1} \sum_{\mathbf{p}_{m-1} \in I(m-1,k-m-1)} S(N+1,\mathbf{p}_{m-1}). \tag{15}$$

Proof. In view of (14) it is clear that (15) is trivially true whenever $k<m$ or $k>2m$, so suppose from the start that $k \leqq 2m \leqq 2k$. It then follows from (12) that for $N \geqq 1$

$$\begin{aligned} S(N,\mathbf{p}_m) &= \sum_{j_1=p_1}^{N+p_1} \sum_{j_2=p_2}^{j_1+p_2} \sum_{j_3=p_3}^{j_2+p_3} \cdots \sum_{j_m=p_m}^{j_{m-1}+p_m} \left(\prod_{i=1}^{m} c_{j_i}^{p_i} b_{j_i}^{1-p_i} \right) \\ &= \sum_{j_1=p_1}^{N+p_1-1} \sum_{j_2=p_2}^{j_1+p_2} \sum_{j_3=p_3}^{j_2+p_3} \cdots \sum_{j_m=p_m}^{j_{m-1}+p_m} \left(\prod_{i=1}^{m} c_{j_i}^{p_i} b_{j_i}^{1-p_i} \right) \\ &\quad + [(c_{N+p_1})^{p_1} (b_{N+p_1})^{1-p_1}] \sum_{j_2=p_2}^{N+p_1+p_2} \sum_{j_3=p_3}^{j_2+p_3} \cdots \sum_{j_m=p_m}^{j_{m-1}+p_m} \left(\prod_{i=2}^{m} c_{j_i}^{p_i} b_{j_i}^{1-p_i} \right) \\ &= S(N-1,\mathbf{p}_m) + [(c_{N+p_1})^{p_1} (b_{N+p_1})^{1-p_1}] S(N+p_1,\mathbf{p}_{m-1}) \end{aligned} \tag{16}$$

(where $\mathbf{p}_{m-1}=(p_2, p_3, \cdots, p_m)$), and thus summing over the index set $I(m, k-m)$ yields

$$(17)\qquad \sum_{\mathbf{p}_m\in I(m,k-m)} S(N,\mathbf{p}_m)=\sum_{\mathbf{p}_m\in I(m,k-m)} S(N-1,\mathbf{p}_m) + \sum_{\mathbf{p}_m\in I(m,k-m)} [(c_{N+p_1})^{p_1}(b_{N+p_1})^{1-p_1}]S(N+p_1,\mathbf{p}_{m-1}).$$

Now $\mathbf{p}_m=(p_1, p_2, \cdots, p_m)=(p_1, \mathbf{p}_{m-1})$ is an ordered m-tuple having $k-m$ ones and $m-(k-m)=2m-k$ zeros. Among the $\binom{m}{k-m}$ possible m-tuples of this type $\binom{m-1}{k-m-1}$ will have $p_1=1$; i.e., $\mathbf{p}_{m-1}$ is an ordered $(m-1)$-tuple having $k-m-1$ ones and $2m-k$ zeros, and consequently $\mathbf{p}_{m-1}\in I(m-1, k-m-1)$. The remaining $\binom{m-1}{k-m}$ ordered m-tuples will have $p_1=0$; and for these remaining ones $\mathbf{p}_{m-1}$ is an ordered $(m-1)$-tuple having $k-m$ ones and $2m-k-1$ zeros, so that $\mathbf{p}_{m-1}\in I(m-1, k-m)$. We may thus rewrite the second sum on the right of (17) as

$$(18)\qquad \sum_{\mathbf{p}_m\in I(m,k-m)} [(c_{N+p_1})^{p_1}(b_{N+p_1})^{1-p_1}]S(N+p_1,\mathbf{p}_{m-1}) = \sum_{\mathbf{p}_{m-1}\in I(m-1,k-m)} b_N S(N,\mathbf{p}_{m-1}) + \sum_{\mathbf{p}_{m-1}\in I(m-1,k-m-1)} c_{N+1}S(N+1,\mathbf{p}_{m-1})$$

which completes the proof for $N\geqq 1$. If $N=0$, the equality in (16) takes the form

$$S(0,\mathbf{p}_m)=[(c_{0+p_1})^{p_1}(b_{0+p_1})^{1-p_1}]S(0+p_1,\mathbf{p}_{m-1})$$

and the conclusion still holds, since $\sum_{\mathbf{p}_m\in I(m,k-m)} S(-1,\mathbf{p}_m)=0$, by (14).

In the recurrence relation (2) the case with constant coefficients

$$(19)\qquad b_n=b\geqq 0,\quad n\geqq 0,\quad c_1=2c>0,\quad c_n=c>0,\quad n\geqq 2$$

turns out to be especially important because it will provide a very nice bound for w_j, leading to convergence for $\sum w_j z^{-j}$. The polynomials generated by this special case are just the monic Chebyshev polynomials of the first kind, orthogonal on the interval $(-b-2\sqrt{c}, -b+2\sqrt{c})$. For these Chebyshev polynomials the sum $\sum_{\mathbf{p}_k\in I(k,q)} S(N,\mathbf{p}_k)$ takes on the following form.

LEMMA 3. *Let b_n, c_n in the recurrence relation* (2) *be the constants prescribed in* (19). *Then for all integers $M\geqq 0$ and all integers j, m satisfying $0\leqq j\leqq M$, $M-j-[(M-j)/2]\leqq m\leqq M-j$,*

$$(20)\qquad \sum_{\mathbf{p}_m\in I(m,M-j-m)} S(j,\mathbf{p}_m)=\binom{M}{2M-j-2m}\binom{2M-j-2m}{M-m} b^{-M+j+2m}c^{M-j-m}.$$

Proof. The proof is by induction on M, using (15) in Lemma 2 (the result (20) was initially guessed by doing an inordinate amount of elementary algebra). It is easy to check that (20) holds for $M=0$, $M=1$. Assume then that it holds for some integer $K>0$, with j, m satisfying $0\leqq j\leqq K$ and $K-j-[(K-j)/2]\leqq m\leqq K-j$. Then by Lemma 2

$$\sum_{\mathbf{p}_m\in I(m,K+1-j-m)} S(j,\mathbf{p}_m)=\sum_{\mathbf{p}_m\in I(m,K+1-j-m)} S(j-1,\mathbf{p}_m) + b\sum_{\mathbf{p}_{m-1}\in I(m-1,K+1-j-m)} S(j,\mathbf{p}_{m-1}) + c\sum_{\mathbf{p}_{m-1}\in I(m-1,K-j-m)} S(j+1,\mathbf{p}_{m-1}).$$

By the induction hypothesis the three sums on the right are, respectively,

$$\sum_{\mathbf{p}_m\in I(m,K+1-j-m)} S(j-1,\mathbf{p}_m)=\sum_{\mathbf{p}_m\in I(m,K-(j-1)-m)} S(j-1,\mathbf{p}_m) = \binom{K}{2K-(j-1)-2m}\binom{2K-(j-1)-2m}{K-m} b^{-K+(j-1)+2m}c^{K-(j-1)-m},$$

$$b \sum_{\mathbf{p}_{m-1} \in I(m-1,\, K+1-j-m)} S(j, \mathbf{p}_{m-1}) = b \sum_{\mathbf{p}_{m-1} \in I(m-1,\, K-j-(m-1))} S(j, \mathbf{p}_{m-1})$$

$$= b \binom{K}{2K-j-2(m-1)} \binom{2K-j-2(m-1)}{K-(m-1)} b^{-K+j+2(m-1)} c^{K-j-(m-1)},$$

$$c \sum_{\mathbf{p}_{m-1} \in I(m-1,\, K-j-m)} S(j+1, \mathbf{p}_{m-1}) = c \sum_{\mathbf{p}_{m-1} \in I(m-1,\, K-(j+1)-(m-1))} S(j+1, \mathbf{p}_{m-1})$$

$$= c \binom{K}{2K-(j+1)-2(m-1)}$$

$$\cdot \binom{2K-(j+1)-2(m-1)}{K-(m-1)} b^{-K+(j+1)+2(m-1)} c^{K-(j+1)-(m-1)},$$

and thus

$$\sum_{\mathbf{p}_m \in I(m,\, K+1-j-m)} S(j, \mathbf{p}_m)$$

$$= \left[\binom{K}{2K+1-j-2m} \binom{2K+1-j-2m}{K-m} + \binom{K}{2K+2-j-2m} \binom{2K+2-j-2m}{K+1-m} \right.$$

$$\left. + \binom{K}{2K+1-j-2m} \binom{2K+1-j-2m}{K+1-m} \right] b^{-(K+1)+j+2m} c^{(K+1)-j-m}.$$

But the first and third terms in the square brackets yield

$$\binom{K}{2K+1-j-2m} \binom{2K+1-j-2m}{K-m} + \binom{K}{2K+1-j-2m} \binom{2K+1-j-2m}{K+1-m}$$

$$= \binom{K}{2K+1-j-2m} \binom{2K+2-j-2m}{K+1-m}$$

and

$$\binom{K}{2K+1-j-2m} \binom{2K+2-j-2m}{K+1-m} + \binom{K}{2K+2-j-2m} \binom{2K+2-j-2m}{K+1-m}$$

$$= \binom{K+1}{2K+2-j-2m} \binom{2K+2-j-2m}{K+1-m},$$

and consequently (20) holds for $M = K+1$.

We are now in position to prove the validity of our conjecture about $\gamma(j, M)$ and then to prove a convergence theorem for $\sum w_j z^{-j}$.

THEOREM 2. *For each integer $M \geqq 0$ and all integers j satisfying $0 \leqq j \leqq M$*

$$\gamma(j, M) = (-1)^{j+M} \sum_{k=M-j-[(M-j)/2]}^{M-j} \sum_{\mathbf{p}_k \in I(k,\, M-j-k)} S(j, \mathbf{p}_k). \tag{21}$$

Proof. We have already seen in (13) that the formula is valid for $\gamma(0, 0), \cdots, \gamma(1, 2)$ and $\gamma(2, 2)$. So assume as induction hypothesis that for some $K > 0$ and all j satisfying $0 \leqq j \leqq K$, (21) is valid for $\gamma(j, K)$ and consider $\gamma(j, K+1)$. From (10) in Lemma 1 and the induction hypothesis we have

$$\begin{aligned}\gamma(j, K+1) &= \gamma(j-1, K) - b_j\gamma(j, K) + c_{j+1}\gamma(j+1, K)\\ &= (-1)^{j-1+K} \sum_{k=K-(j-1)-[(K-(j-1))/2]}^{K-(j-1)} \sum_{\mathbf{p}_k \in I(k, K-(j-1)-k)} S(j-1, \mathbf{p}_k)\\ &\quad - b_j(-1)^{j+K} \sum_{k=K-j-[(K-j)/2]}^{K-j} \sum_{\mathbf{p}_k \in I(k, K-j-k)} S(j, \mathbf{p}_k)\\ &\quad + c_{j+1}(-1)^{j+1+K} \sum_{k=K-(j+1)-[(K-(j+1))/2]}^{K-(j+1)} \sum_{\mathbf{p}_k \in I(k, K-(j+1)-k)} S(j+1, \mathbf{p}_k).\end{aligned} \tag{22}$$

A change of index in the first double sum on the right of (22) converts it to

$$\sum_{k=K-j-[(K-(j-1))/2]}^{K-j} \sum_{\mathbf{p}_{k+1} \in I(k+1, K-j-k)} S(j-1, \mathbf{p}_{k+1}).$$

By using (14) and the fact that the definition of greatest integer function implies

$$K-j-[(K-(j-1))/2] = K-(j+1)-[(K-(j+1))/2] \leqq K-j-[(K-j)/2]$$

we can start all three double sums on the right of (22) at $k = K-j-[(K-(j-1))/2]$, and end all three at $k = K-j$, so that (22) becomes

$$\begin{aligned}\gamma(j, K+1) = (-1)^{j+K+1} \sum_{k=K-j-[(K-(j-1))/2]}^{K-j} \Bigg[&\sum_{\mathbf{p}_{k+1} \in I(k+1, K-j-k)} S(j-1, \mathbf{p}_{k+1})\\ &+ b_j \sum_{\mathbf{p}_k \in I(k, K-j-k)} S(j, \mathbf{p}_k)\\ &+ c_{j+1} \sum_{\mathbf{p}_k \in I(k, K-(j+1)-k)} S(j+1, \mathbf{p}_k) \Bigg].\end{aligned} \tag{23}$$

But this result, by Lemma 2, is

$$\gamma(j, K+1) = (-1)^{j+K+1} \sum_{k=K-j-[(K-(j-1))/2]}^{K-j} \sum_{\mathbf{p}_{k+1} \in I(k+1, K-j-k)} S(j, \mathbf{p}_{k+1}),$$

or, letting $k = i-1$,

$$\gamma(j, K+1) = (-1)^{j+K+1} \sum_{i=K-j+1-[(K-j+1)/2]}^{K-j} \sum_{\mathbf{p}_i \in I(i, K+1-j-i)} S(j, \mathbf{p}_i),$$

valid for $0 \leqq j \leqq K$. The case $j = K+1$ obviously yields $\gamma(K+1, K+1) = 1$, and thus (21) holds for $K+1$ whenever it holds for K.

The sought-for formula for w_{n+1} in terms of b_n and c_n is now available from (21).

COROLLARY 1. *The coefficient w_{n+1} for the Laurent series* (4) *is given by*

$$w_{n+1} = \gamma(0, n) = (-1)^n \sum_{k=n-[n/2]}^{n} \sum_{\mathbf{p}_k \in I(k, n-k)} S(0, \mathbf{p}_k), \qquad n \geqq 0. \tag{24}$$

Applying Lemma 3 to (24) yields the following.

COROLLARY 2. *If b_n and c_n have the constant form* (19), (*i.e., the polynomials are the Chebyshev polynomials of the first kind*) *then denoting the coefficient in this special case by $\gamma_T(0, n)$ we have*

$$\gamma_T(0, n) = (-1)^n \sum_{k=n-[n/2]}^{n} \binom{n}{2n-2k} \binom{2n-2k}{n-k} b^{-n+2k} c^{n-k}. \tag{25}$$

3. Convergence of the Laurent series for $w(z)$.

THEOREM 3. *Let the sequence $\{w_j\}$ be generated by* (8), *and suppose that there exist constants $b \geqq 0$, $c > 0$ such that in* (2) $|b_k| \leqq b$ *for* $k \geqq 0$; $|c_1| \leqq 2c$, $|c_k| \leqq c$ *for* $k \geqq 2$.

Then

$$|w_{n+1}| \leqq |\gamma_T(0, n)| = \sum_{j=0}^{[n/2]} \binom{n}{2j}\binom{2j}{j} b^{n-2j} c^j. \tag{26}$$

Proof. With these bounds on b_k and c_k it is clear from the definition of $S(N, \mathbf{p}_k)$ in (12) and the conclusion (20) in Lemma 3 that

$$\sum_{\mathbf{p}_k \in I(k, n-j-k)} |S(j, \mathbf{p}_j)| \leqq \binom{n}{2n-j-2k}\binom{2n-j-2k}{n-k} b^{-n+j+2k} c^{n-j-k}.$$

Thus (24) and (25) imply $|w_{n+1}| \leqq |\gamma_T(0, n)|$. The change of index $j = n - k$ in (25) yields the sum on the right of (26).

THEOREM 4. *Let the sequence $\{w_j\}$ be generated by* (8), *and suppose that there exist constants $b \geqq 0$, $c > 0$ such that in* (2) $|b_k| \leqq b$ *for* $k \geqq 0$; $|c_1| \leqq 2c$, $|c_k| \leqq c$ *for* $k \geqq 2$. *Furthermore without loss of generality suppose* $b \leqq 2\sqrt{c}$. *Then for all* $n \geqq 0$

$$|w_{n+1}| \leqq \binom{2n}{n}(\sqrt{c})^n, \qquad n \geqq 0. \tag{27}$$

Proof. The hypotheses imply that (26) holds and that $b^{n-2j}c^j \leqq 2^{n-2j}(\sqrt{c})^n$, so from (26) and a combinatorial identity [6, p. 72] there follows

$$|w_{n+1}| \leqq (\sqrt{c})^n \sum_{j=0}^{[n/2]} \binom{n}{2j}\binom{2j}{j} 2^{n-2j} = (\sqrt{c})^n \binom{2n}{n}. \tag{28}$$

Remark. Polynomials generated by the recursion formula (2) when $b_n = 0$ for all $n \geqq 0$ are called symmetric polynomials. In this special case it is clear from (26) that

$$|w_{2n+1}| \leqq \binom{2n}{n} c^n, \qquad w_{2n+2} = 0, \quad n \geqq 0, \tag{29}$$

assuming the boundedness condition stated there for $\{c_n\}$.

We come finally to our main result.

THEOREM 5. *Let $\{p_n\}$ be a sequence of monic polynomials generated by the recurrence relation*

$$p_0 = 1,$$
$$p_1 = z + b_0,$$
$$p_{n+1} = (z + b_n)p_n - c_n p_{n-1}, \qquad n \geqq 1,$$

having complex coefficients b_n, c_n, with $c_n \neq 0$. Suppose there is a constant $c > 0$ such that $|b_n| \leqq 2\sqrt{c}$ for all $n \geqq 0$; $|c_1| \leqq 2c$, $|c_n| \leqq c$ for $n \geqq 2$. Then the series

$$w(z) = \sum_{j=1}^{\infty} w_j z^{-j}$$

with coefficients given by (8) (*or equivalently by* (24)) *converges for all z outside the circle $|z| = 4\sqrt{c}$. Furthermore, on any simple closed contour C having $|z| = 4\sqrt{c}$ in its interior,*

$$(1/2\pi i)\int_C p_n(z)p_m(z)w(z)\,dz = \begin{cases} 0, & m \neq n, \\ 1, & m = n = 0, \\ \prod_{i=1}^{n} c_i, & m = n \geqq 1. \end{cases}$$

Proof. Because of (28) the series $\sum w_j z^{-j}$ converges at least for $|z| > 4\sqrt{c}$ (in the symmetric case it follows from (29) that the convergence holds at least for $|z| > 2\sqrt{c}$). The orthogonality property now follows from Theorem 1.

Remark. A number of examples were helpful in gaining some insight about the kind of conditions to impose on $\{b_n\}$ and $\{c_n\}$ in order to guarantee convergence of the series $\sum w_j z^{-j}$. In addition to the already mentioned monic Chebyshev polynomials, which suggested bounds for the $\{b_n\}$ and $\{c_n\}$, the monic Laguerre polynomials with parameter $\alpha = 0$ show what can happen without a boundedness condition. In this case $b_n = -2n - 1$ for $n \geqq 0$, $c_n = n^2$, $n \geqq 1$. Here the hypotheses of Theorem 5 are clearly not satisfied; in fact we find that $w_j = j!$ for $j \geqq 1$, so $\sum w_j z^{-j}$ diverges for all z.

Acknowledgment. It is appropriate to express here sincere appreciation to Marvin B. Sledd, who first interested me in orthogonal polynomials.

REFERENCES

[1] R. ASKEY AND M. ISMAIL, *Recurrence relations, continued fractions and orthogonal polynomials*, Mem. Amer. Math. Soc., 49 (1984), pp. 8-11.

[2] T. S. CHIHARA, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[3] JA. L. GERONIMUS, *Orthogonal polynomials*, Amer. Math. Soc. Transl. Ser. 2, 108 (1977), pp. 37-130.

[4] E. GROSSWALD, *Bessel polynomials*, Lecture Notes in Mathematics 698, Springer-Verlag, New York, 1978.

[5] P. NEVAI, *Orthogonal polynomials*, Mem. Amer. Math. Soc., 18 (1979), pp. 45ff.

[6] J. RIORDAN, *Combinatorial Identities*, Robert E. Krieger, Huntington, NY, 1979.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

014

THE LINEARIZATION OF THE PRODUCT OF TWO ZONAL POLYNOMIALS*

HOWARD B. KUSHNER†

Abstract. The linearization coefficients $g_{\mu\rho}^{\tau}$ are defined by

$$C_{\mu}(V)C_{\rho}(V)=\sum_{\tau} g_{\mu\rho}^{\tau}C_{\tau}(V)$$

where $C_{\tau}(V)$ is the zonal polynomial corresponding to the partition τ and V is a positive definite matrix. A formula for $g_{(m)\rho}^{\tau}$ is proved, and the partitions τ for which $g_{(m)\rho}^{\tau}\neq 0$ are characterized. An alternative computation of $C_{\tau}(I_k)$ is presented.

Key words. zonal polynomials, linearization of products, partitions

AMS(MOS) subject classifications. 33A75, 62H99

1. Introduction and notation.

Introduction. Zonal polynomials are polynomials of a real, symmetric $k\times k$ matrix V. In the general theory of harmonic analysis, the polynomials are called zonal spherical polynomials of the homogeneous space $\mathrm{GL}(k)/O(k)$, where $\mathrm{GL}(k)$ is the general linear group of invertible real $k\times k$ matrices and $O(k)$ is the orthogonal group of $k\times k$ matrices. They were introduced into multivariate statistical theory by A. T. James, in order to represent certain probability density functions in series form. For example, when V is distributed according to the Wishart distribution, $W_k(N,\Sigma)$—the distribution of $V=X'X$ where the rows of the $N\times k$ matrix X are independently and identically distributed according to the multivariate normal law, $N(0,\Sigma)$—an important problem is to determine the density function of the eigenvalues of V. By introducing the zonal polynomials $\{C_{\tau}(V)\}$—the subscript τ is a partition of t—James (1960) elegantly expressed this density using a bilinear sum of zonal polynomials.

The problem of expressing the product of two zonal polynomials as a linear combination of other zonal polynomials

$$C_{\mu}(V)C_{\rho}(V)=\sum_{\tau} g_{\mu\rho}^{\tau}C_{\tau}(V) \tag{1.1}$$

arose in related statistical problems considered by Constantine (1966), Hayakawa (1967), and Khatri and Pillai (1968), who required the coefficients $g_{\mu\rho}^{\tau}$ in order to compute integrals of products of zonal polynomials. The computation of $g_{\mu\rho}^{\tau}$ is the subject of this paper. Tables for these coefficients were computed by the above authors, but a general formula for them is still unknown. The case $\mu=(1)$ was solved in Kushner (1985). In this paper, the case $\mu=(m)$,

$$C_{(m)}(V)C_{\rho}(V)=\sum_{\tau} g_{(m)\rho}^{\tau}C_{\tau}(V) \tag{1.2}$$

is solved; a simple formula for the coefficients $g_{(m)\rho}^{\tau}$ is given and the partitions τ appearing in (1.2) are characterized. Also, the value of $C_{\tau}(I_k)$ is computed. It appears that the only other way of computing this value is that of James (1961) and Constantine (1963). Kikuchi (1981) has noted that it is not included in Farrell's (1976) treatment of zonal polynomials nor is it in the more recent treatments of Saw (1977), Kates (1981), or Takemura (1984). The present computation is in the spirit of Kushner, Lebow, and Meisner (1981) and thus supplies a missing link in that treatment of zonal polynomials.

* Received by the editors September 23, 1985; accepted for publication May 19, 1987.
† The Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York 10962.

In the mathematical literature, (1.1) is called the linearization of the product of two zonal polynomials; the "g-coefficients" are called the linearization coefficients. Problems of this sort, involving polynomials of one variable, are described in Askey (1975). Hylleraas (1962), for example, formulated a linearization problem in order to compute the integral of a product of Jacobi polynomials. Thus, analytic problems—the computation of integrals—have motivated linearization problems involving zonal polynomials and polynomials of one variable. Nevertheless, more directly related to our problem (1.1) is the better known algebraic problem of linearizing the product of two Schur polynomials (MacDonald (1979)),

$$s_\mu s_\rho = \sum_\tau c_{\mu\rho}^\tau s_\tau. \tag{1.3}$$

The linearization coefficients $c_{\mu\rho}^\tau$ arising in this problem are found by the Littlewood–Richardson rule—a combinatorial algorithm rather than an analytic formula—and are nonnegative integers which give the multiplicity of the representation $\langle\tau\rangle$ of the general linear group in the direct product $\langle\mu\rangle\times\langle\rho\rangle$, as well as the multiplicity of the representation $[\tau]$ of the symmetric group in the direct product $[\mu]\times[\rho]$ (Robinson (1961)). By applying the method of this paper to the case $\mu=(m)$ of linearization problem (1.3),

$$s_{(m)} s_\rho = \sum_\tau c_{(m)\rho}^\tau s_\tau \tag{1.4}$$

one can prove that the partitions τ appearing in (1.4) are precisely those that appear in (1.2) and that $c_{(m)\rho}^\tau = 1$ if τ appears in (1.4). Stanley (1986) has recently found a combinatorial formula for the linearization coefficients (for the case $\mu=(m)$) for a wide class of polynomials, which includes the zonal and Schur polynomials.

Notation.

(A) Partitions.

(A.1) A partition $\tau=(t_1, t_2, \cdots)$ is a finite sequence of nonnegative integers t_i satisfying $t_i \geqq t_{i+1}$. The partition τ is denoted by $\tau=(t_i)$ or by $\tau=[n_i]$, where $t_i - t_{i+1} = n_i$. If $t_i \neq 0$, t_i is called a part of τ. $p=l(\tau)$, the length of τ, is equal to the number of parts of τ. $t_p \neq 0$, but $t_{p+1}=0$. $t=\Sigma t_i$, the norm of τ.

(A.2) $\tau[j]$ is equal to the truncated partition of at most j parts given by $\tau[j]=[n_1, n_2, \cdots, n_j]=(t_1-t_{j+1}, t_2-t_{j+1}, \cdots, t_j-t_{j+1})$.

(A.3) If $\tau=(t_1, t_2, \cdots, t_j)=[n_1, n_2, \cdots, n_j]$ is a partition, the symbol $\tau-l$ is defined by the sequence $\tau-l=(t_1-l, t_2-l, \cdots, t_j-l)=[n_1, n_2, \cdots, n_j-l]$ and therefore the symbol $\tau[j]-l$ by $\tau[j]-l=(t_1-t_{j+1}-l, t_2-t_{j+1}-l, \cdots, t_j-t_{j+1}-l)=[n_1, n_2, \cdots, n_j-l]$. If $t_j \geqq l$, then $\tau-l$ is also a partition. If $t_j < l$, any expression such as $a_{\tau-l}$ will, in this paper, equal zero.

Note that the definition of $\tau-l$ distinguishes between partitions that differ by a string of zeros at the end, e.g., if $\tau_1=(4,2)$ and $\tau_2=(4,2,0)$, then

$$\tau_1-1=(3,1), \qquad \text{a partition}$$

but

$$\tau_2-1=(3,1,-1), \quad \text{a sequence.}$$

(B) Matrices.

(B.1) V is a symmetric $k\times k$ matrix. Usually $V>0$, i.e., V is positive definite. X is a $k\times k$ matrix.

(B.2) V_i is the upper left $i\times i$ matrix of V; $V_k=V$; X_i is the upper left $i\times i$ matrix of X; $X_k=X$. Sometimes (B.2) will not be in effect: X_1, X_2 will denote matrices of a type specified.

(B.3) An upper (right) triangular matrix $T=(t_{ij})$ is one that satisfies $t_{ij}=0$, $i>j$. A lower (left) triangular matrix $S=(s_{ij})$ satisfies $s_{ij}=0$, $i<j$. Upper triangular (respectively, lower triangular) always means upper right triangular (respectively, lower left triangular).

(C) Operators.

(C.1) $D_V=|((1+\delta_{ij})(\partial/\partial v_{ij}))|$ is an operator formed from the determinant of a symmetric $k\times k$ matrix of partial differential operators.

(C.2) $D_X=|(\partial/\partial x_{ij})|$ is an operator formed from the determinant of a $k\times k$ matrix of partial differential operators.

(C.3) D_i is the upper left $i\times i$ determinant of D_V; $D_k=D_V$.

(C.4) If f is a function of V and L is an operator on f, $[Lf](V=0)$ means the evaluation of Lf at the matrix $V=0$. Sometimes this will be written as $L[f](V=0)$. Similarly, if g is a function of X and L is an operator on g, then $[Lg](X)$ means the evaluation of Lg at the matrix X.

(C.5) $L_\tau=\prod_{i=1}^k D_i^{n_i}$, where $\tau=[n_i]$. If $P(V)=\sum_\tau A_\tau C_\tau(V)$, then $2^t t!A_\tau=[L_\tau P](V=0)$ (Kushner and Meisner (1984)).

(D) Eigenfunctions and eigenvalues.

(D.1) If $\tau=(t_i)=[n_i]$ and X is a $k\times k$ matrix, $\Phi_\tau(X)=\prod_{i=1}^k |X_i|^{n_i}$. When $X=V$, a symmetric matrix, $\Phi_\tau(V)=\prod_{i=1}^k |V_i|^{n_i}$ was called a "prototype" polynomial in Kushner, Lebow, and Meisner (1981) and the "power function" in Terras (1985).

$$|V_m|D_m\Phi_\tau=\sigma_m(\tau)\Phi_\tau \quad \text{if } l(\tau)\leqq m,$$

where

$$\sigma_m(\tau)=\prod_{i=1}^m (m-i+2t_i) \quad \text{(Maass (1971, p. 83))}.$$

(D.2) The expectation operator E_n in the Wishart distribution $W_k(n,\Sigma)$ is defined by

$$[E_n f](\Sigma)=C_n|\Sigma|^{-k/2}\int_{V>0} f(V)|V|^{(n-k-1)/2}\exp\left(-\left(\frac{1}{2}\right)\operatorname{tr} V\Sigma^{-1}\right)dV$$

where $C_n^{-1}=2^{(kn)/2}\pi^{k(k-1)/4}\prod_{i=1}^k \Gamma((n-i+1)/2)$. The prototype polynomial, Φ_τ, is an eigenfunction of all the expectation operators:

$$E_n\Phi_\tau=2^t(n/2)_\tau\Phi_\tau$$

where $(a)_\tau=(a)_{t_1}(a-\frac{1}{2})_{t_2}\cdots(a-(k-1)/2)_{t_k}$ is Constantine's generalized hypergeometric symbol (James (1964)).

(E) Miscellaneous.

(E.1) $O(k)$ is the group of orthogonal $k\times k$ matrices. dH is the normalized Haar measure on $O(k)$. All integrals, $\int f(H)\,dH$, are over the full orthogonal group. $f(V)$ is called orthogonally invariant if $f(H'VH)=f(V)$, $H\in O(k)$.

In integrals of the form $\int f(X)\,dX$, where X ranges over the space of $l\times q$ matrices, dX is Lebesgue measure normalized so that

$$\int \exp\left(-\left(\frac{1}{2}\right)\operatorname{tr} X'X\right)dX=(2\pi)^{lq/2}.$$

(E.2) If $a-b\geqq -1$ is an integer, the factorial symbol $[a,b]$ is defined by

$$[a,b]=a(a-1)\cdots b \quad \text{if } a\geqq b,$$

$$[a,b]=1 \quad \text{if } a-b=-1,$$

$$\text{i.e.,}\quad [a,b]=\Gamma(a+1)/\Gamma(b) \quad \text{if } b\neq 0,-1,-2,\cdots.$$

If $a-b \geqq -2$ is an even integer, the factorial symbol $[a, b]_2$ is defined by

$$[a, b]_2 = a(a-2)\cdots b \qquad \text{if } a \geqq b,$$

$$[a, b]_2 = 1 \qquad \text{if } a-b=-2,$$

$$\text{i.e.,} \quad [a, b]_2 = 2^{(a-b+2)/2}\Gamma\left(\frac{a}{2}+1\right)\Big/\Gamma\left(\frac{b}{2}\right) \quad \text{if } b \neq 0, -2, -4, \cdots.$$

(E.3) If (a_i), $1 \leqq i \leqq l$ are vectors in E_n, n-dimensional Euclidean space, then $\{a_i\}$, $1 \leqq i \leqq l$ denotes the vector subspace of E_n spanned by (a_i), $1 \leqq i \leqq n$ (E_n is also used for the expectation operator (D.2), but its meaning will be clear from the context).

2. The product of two zonal polynomials. Let ρ and μ be two partitions of r and m, respectively, whose lengths satisfy $l(\rho) \leqq k$ and $l(\mu) \leqq k$. Then from Kushner and Meisner (1984)

$$C_\mu(V) = a_\mu \int \Phi_\mu(X_1' V X_1)\, e^{-(1/2)\operatorname{tr} X_1' X_1}\, dX_1, \qquad V>0, \tag{2.1a}$$

$$C_\rho(V) = a_\rho \int \Phi_\rho(X_2' V X_2)\, e^{-(1/2)\operatorname{tr} X_2' X_2}\, dX_2, \qquad V>0, \tag{2.1b}$$

where X_1 is a $k \times q_1$ matrix ($q_1 \geqq l(\mu)$), X_2 is a $k \times q_2$ matrix ($q_2 \geqq l(\rho)$), $a_\mu = C_\mu(I_k)/((2\pi)^{q_1 k/2} 2^m (k/2)_\mu)$ and $a_\rho = C_\rho(I_k)/((2\pi)^{q_2 k/2} 2^r (k/2)_\rho)$. It follows that

$$C_\mu(V)C_\rho(V) = a_\mu a_\rho \int \Phi_\mu(X_1' V X_1)\Phi_\rho(X_2' V X_2)\, e^{-(1/2)\operatorname{tr}(X_1' X_1 + X_2' X_2)}\, dX_1\, dX_2. \tag{2.2}$$

Suppose that $l(\rho)+l(\mu) \leqq k$. Choose q_1 and q_2 so that $q_1+q_2=k$. Define a $k \times k$ matrix $X=[X_1; X_2]$ formed by adjoining the matrix X_2 to the matrix X_1. Further, define

$$\bar{V} = (v_{ij}), \qquad q_1+1 \leqq i, j \leqq k, \tag{2.3}$$

$$\bar{\Phi}_\rho(V) = \Phi_\rho(\bar{V}), \tag{2.4}$$

$$G(V) = \Phi_\mu(V)\bar{\Phi}_\rho(V). \tag{2.5}$$

Then (2.2) can be written as

$$C_\mu(V)C_\rho(V) = a_\mu a_\rho \int G(X'VX)\, e^{-(1/2)\operatorname{tr} X'X}\, dX \tag{2.6}$$

where the integration is over the space of $k \times k$ matrices X. Suppose now that $\mu=(m)$, a partition having at most one part, so that $l(\mu) \leqq 1$. Also, suppose $l(\rho) \leqq k-1$. Choose $q_1=1$ and $q_2=k-1$. Define

$$f_m(V) = (1/(2^m m!))\Phi_{(m)}(V) \tag{2.7}$$

and

$$F(V) = f_m(V)\bar{\Phi}_\rho(V). \tag{2.8}$$

Finally, define a function of V by

$$\{m, l, \rho, h\}(V) = |V|^l \int f_m(X_1' V X_1)\bar{\Phi}_\rho(X_2' V X_2) h(X)\, dX \tag{2.9a}$$

$$= |V|^l \int F(X'VX) h(X)\, dX. \tag{2.9b}$$

In (2.9), l is any nonnegative integer and $h(X)$ is a suitable function of X—we will only use rapidly decreasing functions of the form $h(X) = \text{polynomial}\ (X)\ e^{-(1/2)\,\mathrm{tr}\,X'X}$

From (2.6) with $\mu = (m)$ and (2.9) we have that

$$C_{(m)}(V)C_\rho(V) = a_{(m)}a_\rho 2^m m!\{m, 0, \rho, h\}(V) \tag{2.10}$$

where $h(X) \equiv e^{-(1/2)\,\mathrm{tr}\,X'X}$.

In the sequel, $\tau = (t_1, \cdots, t_k) = [n_1, \cdots, n_k]$ denotes an arbitrary partition of t of at most k parts and $\rho = (r_1, \cdots, r_{k-1}) = [m_1, \cdots, m_{k-1}]$ denotes a fixed partition of r at most $k-1$ parts.

3. The L_τ operator. According to Kushner and Meisner (1984) (see (C.5), § 1) the coefficient $g^\tau_{\mu\rho}$ is given by

$$2^t t!\, g^\tau_{\mu\rho} = L_\tau[C_\mu(V)C_\rho(V)](V=0), \qquad l(\tau) \leqq k. \tag{3.1}$$

The goal of this paper is the explicit evaluation of the right side of (3.1), when $\mu = (m)$. In this case, (3.1) can be written as

$$2^t t!\, g^\tau_{(m)\rho} = a_{(m)}a_\rho 2^m m![L_\tau\{m, 0, \rho, h\}](V=0), \qquad l(\tau) \leqq k, \tag{3.2}$$

where the function $\{m, 0, \rho, h\}$ and a_τ are defined in § 2, and L_τ is the operator defined in § 1. In the evaluation of the right side of (3.2), which occupies §§ 5–10, we will utilize two lemmas concerning the operator L_τ. Lemma 1 uses the notation $\tau - i$, $[a, b]$ and $[a, b]_2$ defined in § 1.

LEMMA 1. *If $p(V)$ is an orthogonally invariant polynomial, then*

$$L_\tau[|V|^i p(V)](V=0) = b^i_k(\tau)[L_{\tau-i}p](V=0), \qquad \tau = (t_1, t_2, \cdots, t_k),$$

where

$$b^i_k(\tau) = 2^{ki} \prod_{l=1}^{k} \left[\frac{1}{2}(T_l + k), \frac{1}{2}(T_l + k - 2i + 2)\right], \qquad T_l = 2t_l - l. \tag{3.3}$$

Proof. The zonal polynomials span the orthogonally invariant polynomials (Takemura (1984)). Consequently, $p(V)$ can be expanded into a sum of zonal polynomials

$$p(V) = \sum_\tau A_\tau C_\tau(V) \tag{3.4}$$

where the coefficients A_τ(see (C.5)) are given by

$$2^t t!\, A_\tau = [L_\tau p](V=0). \tag{3.5}$$

Since $|V|^i C_\tau(V) = (C_\tau(I_k)/C_{\tau+i}(I_k))C_{\tau+i}(V)$, we also have

$$\begin{aligned} |V|^i p(V) &= \sum_\tau A_\tau \frac{C_\tau(I_k)}{C_{\tau+i}(I_k)} C_{\tau+i}(V) \\ &= \sum_\tau A_{\tau-i} \frac{C_{\tau-i}(I_k)}{C_\tau(I_k)} C_\tau(V). \end{aligned}$$

Applying (3.5) to the above expansion of the polynomial $|V|^i p(V)$ into a sum of zonal polynomials gives

$$L_\tau[|V|^i p(V)](V=0) = 2^t t!\, A_{\tau-i} \frac{C_{\tau-i}(I_k)}{C_\tau(I_k)}. \tag{3.6}$$

Again using (3.5), but with the partition $\tau - i$ instead of τ, gives

$$2^{t-ki}(t-ki)!\,A_{\tau-i} = [L_{\tau-i}p](V=0). \tag{3.7}$$

Combining (3.6) and (3.7), we obtain

$$L_\tau[|V|^i p(V)](V=0) = \frac{2^t t!}{2^{t-ki}(t-ki)!}\,\frac{C_{\tau-i}(I_k)}{C_\tau(I_k)}[L_{\tau-i}p](V=0). \tag{3.8}$$

We now evaluate the ratio $C_{\tau-i}(I_k)/C_\tau(I_k)$. From James (1964) we have that

$$\frac{2^{2t}t!}{2^{2(t-ki)}(t-ki)!}\,\frac{C_{\tau-i}(I_k)}{C_\tau(I_k)} = \frac{(k/2)_{\tau-i}\prod_{l=1}^k (2t_l-l+k)!}{(k/2)_\tau \prod_{l=1}^k (2t_l-l+k-2i)!}. \tag{3.9}$$

Now

$$\begin{aligned}\frac{(k/2)_\tau}{(k/2)_{\tau-i}} &= \frac{\prod_{l=1}^k ((k-l+1)/2)_{t_l}}{\prod_{l=1}^k ((k-l+1)/2)_{t_l-i}} \\ &= \prod_{l=1}^k \left[\frac{k-l-1}{2}+t_l, \frac{k-l+1}{2}+t_l-i\right] \\ &= \prod_{l=1}^k 2^{-i}[2t_l-l+k-1, 2t_l-l+k-2i+1]_2 \\ &= 2^{-ki}\prod_{l=1}^k [T_l+k-1,\ T_l+k-2i+1]_2.\end{aligned} \tag{3.10}$$

Also,

$$\begin{aligned}\prod_{l=1}^k \frac{(2t_l+k-l)!}{(2t_l-2i+k-l)!} &= \prod_{l=1}^k [2t_l-l+k, 2t_l-l+k-2i+1] \\ &= \prod_{l=1}^k [T_l+k,\ T_l+k-2i+1].\end{aligned} \tag{3.11}$$

Dividing (3.11) by (3.10) evaluates the right side of (3.9) as

$$2^{ki}\prod_{l=1}^k [T_l+k,\ T_l+k-2i+2]_2 = 2^{2ki}\prod_{l=1}^k \left[\frac{1}{2}(T_l+k), \frac{1}{2}(T_l+k-2i+2)\right]$$

and so (3.9) may be written as

$$\frac{2^t t!\,C_{\tau-i}(I_k)}{2^{(t-ki)}(t-ki)!\,C_\tau(I_k)} = 2^{ki}\prod_{l=1}^k \left[\frac{1}{2}(T_l+k), \frac{1}{2}(T_l+k-2i+2)\right]$$

which, by (3.8), is the assertion of the lemma.

LEMMA 2. *If g is a function of the symmetric $k\times k$ matrix V, then*

$$[L_\tau g](V=0) = [L_{\tau[k-1]}G](V_{k-1}=0)$$

where G, a function of the matrix V_{k-1}, is defined by

$$G(V_{k-1}) = [D_k^{n_k} g]\left(V = \begin{pmatrix} V_{k-1} & 0 \\ 0 & 0 \end{pmatrix}\right).$$

Proof. $L_\tau = L_{\tau[k-1]}D_k^{n_k}$ and only D_k involves differentiation with respect to the variables v_{ki}, $1 \leqq i \leqq k$.

4. Matrix factorization; subspace basis. In the sequel, we need to know that, under certain conditions, a square matrix can be factored into various kinds of triangular

matrices and also that a subspace has a "trapezoidal basis." Here, we prove these results. In this section X_i, S_i, T_i denote the $i \times i$ upper left matrix of X, S, T, respectively (see (B) of § 1). Lemma 3, below, is immediately applied in § 5; the other results in this section are not used until § 8.

LEMMA 3. *If the upper left $i \times i$ determinants of the $k \times k$ matrix X satisfy*

$$|X_i| \neq 0, \qquad 1 \leqq i \leqq k,$$

then there exist an upper triangular $k \times k$ matrix, T, and a lower triangular $k \times k$ matrix S, such that

$$X = ST. \tag{4.1}$$

S can be uniquely chosen so that its diagonal entries are all equal to 1.

Proof. The decomposition (4.1) is the well-known Gauss decomposition (Naimark and Stern (1982, p. 249)).

To formulate the next lemma, call an $l \times n$ matrix $A = (a_{ij})$ a "trapezoidal matrix" if, for $1 \leqq i \leqq l$, $a_{ij} = 0$, $n - l + i + 1 \leqq j \leqq n$. (If $l = n$, the matrix A is a lower left triangular matrix.) If $C = \{a_i\}$, $1 \leqq i \leqq l$, where a_i are the row vectors of A, then (a_i) will be called a trapezoidal basis for the vector space C. Related to Lemma 3, but without any of its exceptional cases, is the following result.

LEMMA 4. *Let C be a subspace of E_n, Euclidean n-dimensional space. Suppose that* dim $C = l$. *Then there exists a trapezoidal basis for C.*

Proof. Let G be the $l \times n$ matrix whose row vectors are any basis of C. We must show that there exists an invertible $l \times l$ matrix F such that FG is trapezoidal. If FG is trapezoidal, then

$$f_i \in \{g_{n-l+1+i}, g_{n-l+2+i}, \cdots, g_n\}_\perp = G_i, \qquad 1 \leqq i \leqq l-1$$

where f_i, $1 \leqq i \leqq l$, is the ith row vector of F, g_j, $n - l + 2 \leqq j \leqq n$, is the jth column vector of G, and G_i is the orthogonal complement of the indicated subspace. Vectors f_i and g_i and the subspace G_i all are contained in E_l. G_i is the orthogonal complement of a space having at most $l - i$ independent vectors; hence, dim $G_i \geqq i$. The subspaces G_i satisfy $G_1 \subset G_2 \subset \cdots \subset G_{l-1} \subset G_l = E_l$.

Construct a basis of E_l as follows. Since dim $G_1 \geqq 1$, select $f_1 \neq 0$ from G_1; since dim $G_2 \geqq 2$, select f_2, linearly independent of f_1, from G_2; and so on, until we arrive at a basis, $\{f_i\}$, $1 \leqq i \leqq l$, of E_l with the property: $f_i \in G_i$, $1 \leqq i \leqq l$. The required invertible matrix F is constructed by taking f_i, $1 \leqq i \leqq l$ as its row vectors. (A similar construction was used in Lang (1966, p. 183).)

LEMMA 5. *Any $k \times k$ upper triangular matrix T, all of whose $(i, 1)$-minors are not zero, can be decomposed into $T = AT_1$ where*

$$A = \begin{bmatrix} a_1 & a_1 & 0 & \cdots & & 0 \\ 0 & a_2 & a_2 & \cdots & & 0 \\ 0 & 0 & a_3 & a_3 & \cdots & 0 \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & & & a_{k-1} & a_{k-1} \\ 0 & 0 & \cdots & & 0 & a_k \end{bmatrix}, \quad a_i \neq 0, \quad 1 \leqq i \leqq k,$$

and

$$T_1 = \begin{bmatrix} t_{11} & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & T_0 & \\ 0 & & & \end{bmatrix}.$$

Here T_0 is a $(k-1)\times(k-1)$ upper triangular matrix, and A can be uniquely chosen so that $a_k = 1$.

Proof. Let u_i denote the ith row vector of T_1. Then the ith row vector of AT_1 is

$$a_i(u_i + u_{i+1}) \qquad (u_{k+1} = 0).$$

If t_i denotes the ith row vector of T, define

$$u_l = (-)^k \sum_{i=l}^{k} (-)^i \frac{t_i}{a_i}, \qquad 1 \leqq l \leqq k, \tag{4.2}$$

where the nonzero numbers a_i are to be determined by setting the jth component ($j \geqq 2$) of u_1 equal to zero

$$0 = \sum_{i=1}^{k} (-)^i \frac{t_{ij}}{a_i}, \qquad 2 \leqq j \leqq k.$$

The system

$$0 = \sum_{i=1}^{k} t_{ij} y_i, \qquad 2 \leqq j \leqq k, \tag{4.3}$$

is a system of $k-1$ equations in k unknowns. Let c_i be the co-factor of t_{i1} in the matrix T, i.e.,

$$c_i = (-)^{i+1} C_i,$$

where C_i, the $(i, 1)$-minor of T, is the subdeterminant of T obtained by striking out row i and column 1 from T. Under the conditions of the lemma,

$$y_i = \alpha c_i, \qquad 1 \leqq i \leqq k,$$

α a scalar, exhausts all the solutions of (4.3). Define then $a_i = C_i^{-1}$, and the row vectors of T_1 by (4.2), to obtain the assertion of the lemma.

If R is a $k \times k$ matrix, let $R(j)$ denote the submatrix of R formed from rows 1 to j and columns 2 to $j+1$.

LEMMA 6. *If $|X(j)| \neq 0$, $1 \leqq j \leqq k-1$, and if $|X_l| \neq 0$, $1 \leqq l \leqq k$, then there exist a lower triangular matrix S, matrices A and T_1 of the form given in Lemma 5, such that $X = SAT_1$. The a_i may be chosen so that $a_i = 1$, $1 \leqq i \leqq k$.*

Proof. By Lemma 3,

$$X = ST \tag{4.4}$$

where S and T are lower and upper triangular, respectively, and $s_{ii} = 1$, $i \leqq k$. The plan of the proof is to apply Lemma 5 to the T in (4.4). T is invertible since $|S||T| = |X| \neq 0$. The $(i, 1)$-minors of T are equal to $\pm|T|b_i$, where b_i, $1 \leqq i \leqq k$, are the components of the first row vector of T^{-1}. Let $\bar{b}_l = (b_1, \cdots, b_l)$, an l-dimensional vector. Then from

$$\bar{b}_l T_l' = (1, 0, \cdots, 0)$$

we find

$$b_l = \frac{|T(l-1)|}{|T_l|}. \tag{4.5}$$

The matrices $X(l-1)$ and $T(l-1)$ are related by

$$X(l-1) = S_{l-1} T(l-1)$$

which follows from (4.4). Since $|S_{l-1}| = 1$, we obtain

$$|X(l-1)| = |T(l-1)|.$$

From (4.5), we then find that the $(l, 1)$-minor of T is given by

$$(-)^{l+1}|T|b_l = (-)^{l+1}\left(\prod_{j=l+1}^{k} t_{jj}\right)|X(l-1)|.$$

Certainly no t_{jj} is zero, since $\prod_{j=1}^{k} t_{jj} = |X| \neq 0$. We conclude that none of the $(i, 1)$-minors of T are equal to zero. Lemma 5 then permits us to write

$$T = AT_1.$$

Substituting the above representation of T in (4.4) gives the first assertion of the lemma. Replacing S by SD^{-1} where $D = \text{diag}\,(a_1, a_2, \cdots, a_k)$ gives the second assertion.

5. Operators bi-invariant with respect to matrix multiplication: eigenfunctions, eigenvalues, and an integral. An operator L is bi-invariant with respect to left and right matrix multiplication if it satisfies the chain rule property

$$[LF](X) = [Lf](T_1XT_2). \tag{5.1}$$

In (5.1), f is a polynomial of the $k \times k$ matrix X and F is the polynomial of X defined, for any two fixed matrices T_1 and T_2, by

$$F(X) = f(T_1XT_2).$$

The operator $L = |X|D_X$ is bi-invariant with respect to right and left matrix multiplication. Let T_1 and T_2 be two upper triangular matrices. The polynomial (§ 1) $\Phi_\tau(X)$ satisfies

$$\Phi_\tau(T_1'XT_2) = \Phi_\tau(X)\Phi_\tau(T_1'T_2). \tag{5.2}$$

This property is a generalization of the well-known property

$$\Phi_\tau(T'VT) = \Phi_\tau(V)\Phi_\tau(T'T), \quad T \text{ upper triangular},$$

of the "prototype" polynomials $\Phi_\tau(V)$. We will now use (5.2) to show that $\Phi_\tau(X)$ is an eigenfunction of every bi-invariant operator. The following proof of this assertion is modeled on the proof of the analogous statement in Selberg (1956) or Maass (1971).

THEOREM 1. *The polynomial $\Phi_\tau(X)$ is an eigenfunction of every bi-invariant operator.*

Proof. Let the bi-invariant operator L operate on both sides of (5.2), obtaining, for any two upper triangular matrices,

$$[L\Phi_\tau](T_1'XT_2) = \Phi_\tau(T_1'T_2)[L\Phi_\tau](X).$$

Setting $X = I$, the $k \times k$ identity, in the above equation yields

$$[L\Phi_\tau](T_1'T_2) = \lambda\Phi_\tau(T_1'T_2)$$

where $\lambda = [L\Phi_\tau](I)$. According to Lemma 3, every $k \times k$ matrix X satisfying $|X_i| \neq 0$, $1 \leqq i \leqq k$, can be represented in the form $X = T_1'T_2$, where T_1 and T_2 are upper triangular. Therefore, the eigenfunction property

$$[L\Phi_\tau](X) = \lambda\Phi_\tau(X) \tag{5.3}$$

holds, except possibly at matrices X for which one or more of the determinants $|X_i|$ vanish. It follows that (5.3) also holds at any matrix X for which both sides of the above equation are continuous, proving the theorem.

LEMMA 7. *Suppose $f(X) \geqq 0$ is an eigenfunction of $|X|D_X$ and that f satisfies either*

(5.4a) (a) $$f(XT) = \Phi_\mu(T'T)f(X)$$

where T is upper triangular, or

(5.4b) (b) $$f(SX) = \Phi_\mu(S'S)f(X)$$

where S is lower triangular. Then

(5.5) $$|X|D_X f = \sigma_k(\mu)f$$

where $\sigma_k(\mu)$ (given in § 1) is the eigenvalue in

(5.6) $$|V|D_V\Phi_\mu = \sigma_k(\mu)\Phi_\mu.$$

Proof. We confine ourselves to (a), the upper triangular version of the lemma. Define a function g of the symmetric matrix $V = X'X$ by

$$g(X'X) = \int f(HX)\, dH.$$

Setting $X = T$, an upper triangular matrix, in the above equation and using (5.4a) gives

$$g(T'T) = c\Phi_\mu(T'T)$$

where $c = \int f(H)\, dH$. If $X'X$ is positive definite, an upper triangular T can be found such that $X'X = T'T$. Therefore

$$g(X'X) = c\Phi_\mu(X'X)$$

and

(5.7) $$c\Phi_\mu(X'X) = \int f(HX)\, dH$$

whenever X is invertible. If $c = 0$ in (5.7), then f vanishes identically. For, $\int f(HX)\, dH = 0$ and $f(HX) \geqq 0$ imply that $f(HX) = 0$ for every $H \in O(k)$; in particular, $f(X) = 0$. Assume then that $c \neq 0$.

The operators $|X|D_X$ and $|V|D_V$ act identically on functions of the variable $X'X$. If $F(X) = G(X'X)$, then

(5.8) $$|X|D_X F = |V|D_V G, \qquad V = X'X.$$

Also, by assumption,

(5.9) $$|X|D_X f = \beta f.$$

Operate with the invariant operator $|X|D_X$ on both sides of (5.7), using (5.8) and (5.9). We obtain

$$c\sigma_k(\mu)\Phi_\mu(X'X) = \beta \int f(HX)\, dX = c\beta\Phi_\mu(X'X)$$

from which follows the equality

$$\beta = \sigma_k(\mu),$$

that is,

$$|X|D_X f = \sigma_k(\mu)f.$$

According to Theorem 1, the polynomial $\Phi_\tau(X)$ is an eigenfunction of $|X|D_X$. We now compute its eigenvalue.

THEOREM 2.

$$|X|D_X\Phi_\tau = \sigma_k(\tau/2)\Phi_\tau. \tag{5.10}$$

Proof. Setting $T_1 = I$ in (5.2) shows that $\Phi_\tau(X)$ satisfies the functional equation

$$\begin{aligned}\Phi_\tau(XT) &= \Phi_\tau(T)\Phi_\tau(X)\\ &= \Phi_{\tau/2}(T'T)\Phi_\tau(X), \quad T \text{ upper triangular.}\end{aligned}$$

Now assume that the parts of τ are even integers. Then $\Phi_\tau(X) \geqq 0$. The three conditions of Lemma 7 are now satisfied by the polynomial Φ_τ, so the assertion of the theorem is established for this case. Next, for any partition τ, consider the equation, obtained by evaluating (5.10) at $X = I$

$$[|X|D_X\Phi_\tau](I) = \sigma_k(\tau/2). \tag{5.11}$$

Equation (5.11) has just been proved in the case when the parts of τ are all even. Both sides of (5.11) have an obvious meaning even when t_i, the "parts" of τ are arbitrary real numbers. Indeed, both sides of (5.11) are polynomials in t_i, $1 \leqq i \leqq k$, and, since both sides of (5.11) are equal when all the t_i are even integers, we may deduce the truth of (5.11) for all real t_i. In particular, (5.11) is true for any partition τ. Theorem 2 now follows from Theorem 1 and (5.11).

The remainder of this section is devoted to the evaluation of the integrals (5.14), (5.16), (5.17), below. The evaluation repeatedly uses "integration by parts," unlike the evaluation of the analogous integral, $\int_{V>0} \Phi_\tau(V)\, e^{-(1/2)\operatorname{tr} V}\, dV$, which is directly accomplished by a change of variables, as in Selberg (1956), Constantine (1963), or Maass (1971). The integration by parts formula to be used is

$$fD_Xg - (-)^k gD_Xf = \operatorname{div} B. \tag{5.12}$$

In (5.12), f and g are functions of X, B is a vector of functions of X, and div is the divergence operator. Equation (5.12) follows from a more general result in Maass (1971); a different derivation of (5.12) is in Kushner (1980). If f and g are suitable functions, then

$$\int fD_Xg\, dX = (-)^k \int gD_Xf\, dX \tag{5.13}$$

follows from (5.12). In particular, we may use (5.13) when f is a polynomial in X and g is a polynomial in X times the function $e^{-(1/2)\operatorname{tr} X'X}$.

THEOREM 3. *Define c_τ by the integral*

$$c_\tau = (2\pi)^{-k^2/2} \int \Phi_\tau(X)\, e^{-(1/2)\operatorname{tr} X'X}\, dX. \tag{5.14}$$

Then

$$\begin{aligned}&c_{2\tau} = \prod_{l \leqq i}^{p} [T_l - T_{i+1} - 2,\ T_l - T_i + 1]_2, \qquad p = l(\tau), \quad T_i = 2t_i - i,\\ &c_\tau = 0 \quad \textit{if the } t_i \textit{ are not all even.}\end{aligned} \tag{5.15}$$

Also,

$$\text{(a)} \qquad 2^{r/2}(k/2)_{\tau/2} \int \Phi_\tau(H)\, dH = c_\tau \tag{5.16}$$

where $(a)_\mu$ *is Constantine's generalized hypergeometric symbol defined in* (D.2) *and*

$$(5.17)\qquad (b)\quad (2\pi)^{-k^2/2}\int \Phi_\tau(X)\, e^{-(1/2)\,\mathrm{tr}\, A^{-1}X'X}\, dX = \Phi_{\tau/2}(A)|A|^{k/2}c_\tau, \qquad A>0.$$

Proof. (a) By the substitution $X = HT$, $H \in O(k)$, T an upper triangular matrix with nonnegative diagonal entries, we have

$$dX = 2^k g_k \prod_{i=1}^{k} t_{ii}^{k-i}\, dH\, dT,$$

$$g_k = \pi^{k(k+1)/4} \Big/ \prod_{i=1}^{k} \Gamma(i/2),$$

$$\Phi_\tau(HT) = \Phi_\tau(H)\Phi_\tau(T),$$

so c_τ, as defined in (5.14), is given by

$$(5.18)\qquad c_\tau = (2\pi)^{-k^2/2}2^k g_k \int \Phi_\tau(T)\, e^{-(1/2)\,\mathrm{tr}\, T'T} \prod_{i=1}^{k} t_{ii}^{k-i}\, dT \int \Phi_\tau(H)\, dH.$$

But $\Phi_\tau(T) = \Phi_{\tau/2}(T'T)$, and if the change of variables, $V = T'T$, with $dV = 2^k \prod_{i=1}^{k} t_{ii}^{k+1-i}\, dT$ is used in (5.18), we obtain

$$\begin{aligned} c_\tau &= (2\pi)^{-k^2/2} g_k \int_{V>0} \Phi_{\tau/2}(V)\, e^{-(1/2)\,\mathrm{tr}\, V} \frac{dV}{|V|^{1/2}} \int \Phi_\tau(H)\, dH \\ &= [E_k \Phi_{\tau/2}](I_k) \int \Phi_\tau(H)\, dH \end{aligned} \qquad (5.19)$$

where E_k is the expectation operator in the Wishart $W_k(k, \Sigma)$ distribution. By $\tau/2$ we mean the partition $\tau/2 = (t_i/2)$, $1 \leqq i \leqq k$, and by $\Phi_{\tau/2}(V)$ we of course mean the function

$$\Phi_{\tau/2}(V) = \prod_{i=1}^{k} |V_i|^{t_i/2},$$

which is, in the terminology of Kushner, Lebow, and Meisner (1981), an *EP* function, satisfying $E_k\Phi_{\tau/2} = 2^{t/2}(k/2)_{\tau/2}\Phi_{\tau/2}$ by Constantine (1963). In particular,

$$(5.20)\qquad [E_k\Phi_{\tau/2}](I_k) = 2^{t/2}(k/2)_{\tau/2}.$$

Substituting (5.20) in (5.19) gives

$$c_\tau = 2^{t/2}(k/2)_{\tau/2} \int \Phi_\tau(H)\, dH.$$

(b) Write $A^{-1} = TT'$, T an upper triangular matrix. By the substitution $Y = XT$, we have

$$dY = |T|^k\, dX, \qquad \Phi_\tau(YT^{-1}) = \Phi_\tau(Y)\Phi_\tau(T^{-1})$$

so the integral (5.17) is given by

$$\begin{aligned} &(2\pi)^{-k^2/2} \int \Phi_\tau(X)\, e^{-(1/2)\,\mathrm{tr}\, A^{-1}X'X}\, dX \\ &\quad = \Phi_\tau(T^{-1})|T|^{-k} \int \Phi_\tau(Y)\, e^{-(1/2)\,\mathrm{tr}\, Y'Y}\, dY \\ &\quad = \Phi_{\tau/2}((T^{-1})'T^{-1})|TT'|^{-k/2} \int \Phi_\tau(Y)\, e^{-(1/2)\,\mathrm{tr}\, Y'Y}\, dY \\ &\quad = \Phi_{\tau/2}(A)|A|^{k/2} \int \Phi_\tau(Y)\, e^{-(1/2)\,\mathrm{tr}\, Y'Y}\, dY \\ &\quad = \Phi_{\tau/2}(A)|A|^{k/2}c_\tau. \end{aligned}$$

Next, we prove that if the t_i are not all even, then the integral (5.16) is zero. Let $K = \operatorname{diag}(1, 1, \cdots, -1, 1, \cdots, 1)$ with the -1 lying in the lth place. Then $K \in O(k)$ is upper triangular and $\Phi_\tau(HK) = \Phi_\tau(H)\Phi_\tau(K) = (-)^{t_l}\Phi_\tau(H)$. Using the substitution $H \to HK$ in (5.16), we obtain

$$c_\tau = (-)^{t_l} c_\tau \tag{5.21}$$

or $c_\tau = 0$ if any t_l is odd.

Finally, we now evaluate the integral (5.14) when t_i are all even. Let $h(X) = e^{-(1/2)\operatorname{tr} X'X}$, and note that $D_X h = (-)^k |X| h$. Then with $\tau = (t_1, t_2, \cdots, t_k)$,

$$\begin{aligned} \int \Phi_\tau(X) h(X)\, dX &= (-)^k \int \Phi_{\tau-1}(X) D_X h\, dX \\ &= \int [D_X \Phi_{\tau-1}](X) h(X)\, dX \\ &= \sigma_k((\tau-1)/2) \int \Phi_{\tau-2}(X) h(X)\, dX; \end{aligned} \tag{5.22}$$

the last two steps follow from (5.13) and (5.10).

Doing this $n_k/2$ times we obtain

$$\int \Phi_\tau(X) h(X)\, dX = \prod_{j=1}^{n_k/2} \sigma_k((\tau-(2j-1))/2) \int \Phi_{\tau[k-1]}(X) h(X)\, dX \tag{5.23}$$

$$\begin{aligned} &= (2\pi)^{(2k-1)/2} \prod_{j=1}^{n_k/2} \sigma_k((\tau-(2j-1))/2) \\ &\quad \cdot \int \Phi_{\tau[k-1]}(X_{k-1}) h(X_{k-1})\, dX_{k-1} \end{aligned} \tag{5.24}$$

where $(\tau-(2j-1))/2$ is the "partition" whose ith part is $(t_i-(2j-1))/2$ and X_{k-1} is the $(k-1)\times(k-1)$ upper left submatrix of X. If $n_k = 0$, the empty product $\prod_1^0$ may be omitted from (5.24). Again repeating this process, we obtain

$$\int \Phi_\tau(X) h(X)\, dX = (2\pi)^{k^2/2} \prod_{i=1}^{p} \prod_{j=i}^{n_i/2} \sigma_i((\tau[i]-(2j-1))/2), \qquad p = l(\tau). \tag{5.25}$$

Now the lth part of the "partition" $(\tau[i]-(2j-1))/2$ is $(t_l - t_{i+1}-(2j-1))/2$, $1 \leqq l \leqq i$. By (D.1) of § 1, $\sigma_i((\tau[i]-(2j-1)/2)) = \prod_{l=1}^{i}(i-l+t_l-t_{i+1}-(2j-1))$, and (5.25) becomes

$$\begin{aligned} \int \Phi_\tau(X) h(X)\, dX &= (2\pi)^{k^2/2} \prod_{i=1}^{p} \prod_{j=1}^{n_i/2} \prod_{l=1}^{i} (i-l+t_l-t_{i+1}-(2j-1)) \\ &= (2\pi)^{k^2/2} \prod_{i=1}^{p} \prod_{\text{odd } j=1}^{t_i-t_{i+1}} \prod_{l=1}^{i} (i-l+t_l-t_{i+1}-j) \qquad (t_{p+1}=0). \end{aligned} \tag{5.26}$$

Writing 2τ for τ in (5.26) and noting that

$$\prod_{\text{odd } j=1}^{2(t_i-t_{i+1})} (i-l+2t_l-2t_{i+1}-j) = [T_l - T_{i+1} - 2,\ T_l - T_i + 1]_2,$$

allows (5.26) to be written as

$$\int \Phi_{2\tau}(X) h(X)\, dX = (2\pi)^{k^2/2} \prod_{l \leqq i}^{p} [T_l - T_{i+1} - 2,\ T_l - T_i + 1]_2$$

which is formula (5.15).

6. Invariant operators; computation of $C_\tau(I_k)$; lengths of partitions. An operator L invariant with respect to congruence transformations, $V \to T'VT$, is one that obeys the simple chain rule law

$$[Lf_T](V) = [Lf](T'VT), \quad T \text{ any real matrix},$$

where f is a polynomial of the positive definite matrix V and $f_T(V) = f(T'VT)$. More generally, if

$$g(V) = \int F(X'VX)h(X)\,dX, \quad F(V) \text{ a polynomial}, \tag{6.1}$$

then it can be shown that

$$[Lg](V) = \int [LF](X'VX)h(X)\,dX. \tag{6.2}$$

Let $D = D_V$. With $L = |V|^n D^n$, an invariant operator with respect to congruence transformations, (6.2) becomes

$$|V|^n[D^n g](V) = |V|^n \int [D^n F](X'VX)|X|^{2n}h(X)\,dX$$

from which we obtain

$$[D^n g](V) = \int [D^n F](X'VX)|X|^{2n}h(X)\,dX. \tag{6.3}$$

As a first application of (6.3), we now compute the value of $C_\tau(I_k)$. As mentioned in § 1, it appears that the only other computation of $C_\tau(I_k)$ is that of James (1961) and Constantine (1963).

From Kates (1981) or Kushner and Meisner (1984),

$$(2\pi)^{k^2/2}2^t(k/2)_\tau C_\tau(V) = C_\tau(I_k)\int \Phi_\tau(X'VX)\,e^{-(1/2)\,\mathrm{tr}\,X'X}\,dX. \tag{6.4}$$

But from Kushner and Meisner (1984) (see (C.5)),

$$[L_\tau C_\tau](V=0) = 2^t t!. \tag{6.5}$$

Operating with L_τ on (6.4) and using (6.5) yields

$$(2\pi)^{k^2/2}2^{2t}t!(k/2)_\tau = C_\tau(I_k)\left[L_\tau \int \Phi_\tau(X'VX)\,e^{-(1/2)\,\mathrm{tr}\,X'X}\,dX\right](V=0). \tag{6.6}$$

By evaluating the right side of (6.6), we will obtain a formula for $C_\tau(I_k)$. The procedure is to repeatedly apply Lemma 2, Theorem 2, and (6.3).

Using (6.1) and (6.3) with $F(V) = \Phi_\tau(V)$ and $h(X) = e^{-(1/2)\,\mathrm{tr}\,X'X}$, the right side of (6.6) is, by Lemma 2 and Theorem 2,

$$C_\tau(I_k)\sigma_k^{n_k}(\tau)\left[L_{\tau[k-1]}\int \Phi_{\tau[k-1]}(X'_{k-1}V_{k-1}X_{k-1})|X|^{2n_k}\,e^{-(1/2)\,\mathrm{tr}\,X'X}\,dX\right](V_{k-1}=0) \tag{6.7}$$

where

$$\sigma_k^n(\tau) = \prod_{j=0}^{n-1}\sigma_k(\tau - j) \quad \text{if } n \geqq 1 \quad (\tau = (t_1, t_2, \cdots, t_k)),$$

$$\sigma_k^0(\tau) = 1$$

and

$$L_{\tau[k-1]} = \prod_{i=1}^{k-1} D_i^{n_i}. \tag{6.8}$$

Continuing in this fashion enables us to evaluate the right side of (6.6) and so to obtain the equation

$$\begin{aligned}(2\pi)^{k^2/2} 2^{2t} (k/2)_\tau t! = C_\tau(I_k) \prod_{i=1}^{p} \prod_{j=0}^{n_i-1} \sigma_i(\tau[i]-j) \\ \cdot \int \Phi_{2\tau}(X)\, e^{-(1/2)\operatorname{tr} X'X}\, dX, \qquad p = l(\tau).\end{aligned} \tag{6.9}$$

The product in (6.9) is evaluated as was the product in (5.25). We obtain

$$\begin{aligned}\prod_{i=1}^{p} \prod_{j=0}^{n_i-1} \sigma_i(\tau[i]-j) &= \prod_{i=1}^{p} \prod_{j=0}^{n_i-1} \prod_{l=1}^{i} (i-l+2(t_l-t_{i+1})-2j) \\ &= \prod_{i=1}^{p} \prod_{\text{even } j=0}^{2(t_i-t_{i+1})-2} \prod_{l=1}^{i} (i-l+2(t_l-t_{i+1})-j) \\ &= \prod_{l \leqq i}^{p} [T_l - T_{i+1} - 1,\ T_l - T_i + 2]_2.\end{aligned} \tag{6.10}$$

Finally, we use (5.15) and (6.10) to evaluate the right side of (6.9), obtaining

$$\begin{aligned}2^{2t}(k/2)_\tau t! &= C_\tau(I_k) \prod_{l \leqq i}^{p} [T_l - T_{i+1} - 1,\ T_l - T_i + 2]_2 \\ &\quad \cdot [T_l - T_{i+1} - 2,\ T_l - T_i + 1]_2\end{aligned} \tag{6.11}$$

$$= C_\tau(I_k) \prod_{l \leqq i}^{p} [T_l - T_{i+1} - 1,\ T_l - T_i + 1]. \tag{6.12}$$

We now show that the formula (6.12) for $C_\tau(I_k)$ is equivalent to that of James and Constantine. In (6.12), when $T_l - T_{i+1} - 1 < T_l - T_i + 1$, the symbol $[T_l - T_{i+1} - 1,\ T_l - T_i + 1] \equiv 1$. This happens when $2 > T_i - T_{i+1} = 2(t_i - t_{i+1}) + 1$, implying that $t_i = t_{i+1}$. Even in this case, when $i \geqq l$,

$$[T_l - T_{i+1} - 1,\ T_l - T_i + 1] = \frac{(T_l - T_{i+1} - 1)!}{(T_l - T_i)!}. \tag{6.13}$$

Noting that $T_l - T_i > 0$ for $i > l$, it follows from (6.13) that

$$\begin{aligned}\prod_{i=l}^{p} [T_l - T_{i+1} - 1,\ T_l - T_i + 1] &= \prod_{i=l}^{p} \frac{(T_l - T_{i+1} - 1)!}{(T_l - T_i)!} \\ &= (T_l - T_{l+1} - 1)! \prod_{i>l}^{p} \frac{(T_l - T_{i+1} - 1)!}{(T_l - T_i)!} \\ &= (T_l - T_{l+1} - 1)! \prod_{i>l}^{p} \frac{(T_l - T_{i+1} - 1)!}{(T_l - T_i)(T_l - T_i - 1)!} \\ &= (T_l - T_{l+1} - 1)! \frac{(T_l - T_{p+1} - 1)!}{(T_l - T_{l+1} - 1)!} \left[\prod_{i>l}^{p} (T_l - T_i) \right]^{-1} \\ &= \frac{(T_l + p)!}{\prod_{i>l}^{p} (T_l - T_i)} \quad \text{(fixed } l\text{)},\end{aligned} \tag{6.14}$$

which, with (6.12), gives the James-Constantine formula for $C_\tau(I_k)$ (James (1964)),

$$2^{2t}(k/2)_\tau t! = C_\tau(I_k)\frac{\prod_{l=1}^{p}(T_l+p)!}{\prod_{i>l}^{p}(T_l-T_i)}.$$

As a second application of (6.3), we now prove that the lengths of the partitions τ appearing in (1.1) satisfy

$$\max(l(\mu), l(\rho)) \leqq l(\tau) \leqq l(\mu)+l(\rho). \tag{6.15}$$

Let $G(V) \equiv \Phi_\mu(V)\Phi_\rho(\bar{V})$ as in (2.5). Then G is a function of the variables in V_j where $j = l(\mu)+l(\rho)$, implying $D_iG \equiv 0$, if $i > j$. Hence, by (6.3) and (2.6) (thinking of the k in (2.6) as i), $D_iC_\mu(V)C_\rho(V) \equiv 0$, if $i > l(\mu)+l(\rho)$. Consequently, if $l(\tau) > l(\mu)+l(\rho)$, then $0 = [L_\tau C_\mu(V)C_\rho(V)](V=0) = 2^t t! g^\tau_{\mu\rho}$, from (3.1). This proves the right side of (6.15). To prove the left side note that if $i < l(\mu)$, then $C_\mu(V_i) = 0$, which implies that

$$C_\mu(V_i)C_\rho(V_i) = \sum_\tau g^\tau_{\mu\rho}C_\tau(V_i) \equiv 0$$

or

$$g^\tau_{\mu\rho} = 0 \quad \text{if } l(\tau) \leqq i < l(\mu).$$

In calculating linearization coefficients via (3.1), it therefore suffices to take any $k \geqq l(\mu)+l(\rho)$.

7. *g*-coefficients: characterization of partitions; integral representation. In this section, we characterize the partitions τ appearing in (1.2) whose g-coefficient is not zero. Also, an integral formula for these g-coefficients is derived. The positivity of the coefficients is an immediate consequence of the integral formula, which is finally evaluated in § 10.

THEOREM 4. *Suppose* $\tau = (t_j)$, $1 \leqq j \leqq p+1$, $\rho = (r_j)$, $1 \leqq j \leqq p$ *and* $r_{p+1} = 0$. *If*

$$t_j \geqq r_j \geqq t_{j+1}, \qquad 1 \leqq j \leqq p \tag{7.1a}$$

and

$$t = r + m, \tag{7.1b}$$

then

$$2^t t! g^\tau_{(m)\rho} = C \int \Phi_\mu(X)\Phi_\nu(X^*)\, e^{-(1/2)\operatorname{tr} X'X}\, dX \tag{7.2a}$$

where

$$\mu = 2[t_j - r_j], \qquad 1 \leqq j \leqq p+1, \tag{7.2b}$$

$$\nu = 2[r_j - t_{j+1}], \quad 1 \leqq j \leqq p, \quad p = l(\rho), \tag{7.2c}$$

$$C = m! a_{(m)}a_\rho 2^t \prod_{l \leqq j}^{p} \left[\frac{1}{2}(R_l - R_{j+1} - 1), \frac{1}{2}(R_l - T_{j+1} + 1)\right] \cdot \left[\frac{1}{2}(T_l - T_{j+1} - 1), \frac{1}{2}(T_l - R_j + 2)\right], \tag{7.2d}$$

$$X = (x_{ij}), \qquad 1 \leqq i, j \leqq p+1, \tag{7.2e}$$

$$X^* = (x_{ij}), \quad 1 \leqq i \leqq p, \quad 2 \leqq j \leqq p+1, \quad \textit{the upper right } p \times p \textit{ part of } X. \tag{7.2f}$$

In (7.2d), $T_l = 2t_l - l$, $R_l = 2r_l - l$. a_τ *is defined in* § 2.

If τ does not satisfy (7.1), *then* $g^{\tau}_{(m)\rho} = 0$.

Proof. The proof starts with the expression for $g^{\tau}_{(m)\rho}$ given in (3.2). Let $k = p+1$. From (6.15), $l(\tau) \leqq k$. The first goal is to calculate the effect of D^n_k on the function $\{m, 0, \rho, h\}$. The procedure is similar to, but more complicated than, that used in § 6.

With $F(V)$ defined by (2.8), the function g defined in (6.1) is just $g(V) \equiv \{m, 0, \rho, h\}$. From (6.3), we therefore have

$$D^n_k\{m, 0, \rho, h\} = \int [D^n_k F](X'VX)|X|^{2n} h(X)\, dX. \tag{7.3}$$

We now calculate $D^n_k F$. By the Laplace expansion of a determinant, we can write the operator D_k as

$$D_k = \left(2\frac{\partial}{\partial v_{11}}\right)\bar{D}_{k-1} + \text{other terms}$$

where $\bar{D}_{k-1}$ is the $(k-1)\times(k-1)$ minor of $2(\partial/\partial v_{11})$ in D_k. The "other terms" all contain a $\partial/\partial v_{1i}$, $2 \leqq i \leqq k$. Therefore, $D_k F = D_k[f_m(V)\bar{\Phi}_\rho(V)] = (2(\partial/\partial v_{11}))f_m \bar{D}_{k-1}\bar{\Phi}_\rho$. For any n,

$$D^n_k F = \left(2\frac{\partial}{\partial v_{11}}\right)^n f_m \bar{D}^n_{k-1}\bar{\Phi}_\rho. \tag{7.4}$$

We now simplify the right side of (7.4). First, note that $(2(\partial/\partial v_{11}))f_m = f_{m-1}$. For any n,

$$\left(2\frac{\partial}{\partial v_{11}}\right)^n f_m = f_{m-n}, \qquad f_i \equiv 0 \quad \text{if } i < 0. \tag{7.5}$$

Second, from § 1, we have that $|V_{k-1}|D_{k-1}\Phi_\rho = \sigma_{k-1}(\rho)\Phi_\rho$ or $D_{k-1}\Phi_\rho = \sigma_{k-1}(\rho)\Phi_{\rho-1}$. With $\rho = (r_i) = [m_i]$, the symbol $\rho - 1 = (r_j - 1)$, $1 \leqq j \leqq k-1$. Similarly, for any n,

$$|V_{k-1}|^n D^n_{k-1}\Phi_\rho = \sigma^n_{k-1}(\rho)\Phi_\rho$$

or

$$D^n_{k-1}\Phi_\rho = \sigma^n_{k-1}(\rho)\Phi_{\rho-n}, \qquad \Phi_{\rho-n} \equiv 0 \quad \text{if } m_{k-1} < n, \tag{7.6}$$

and

$$\bar{D}^n_{k-1}\bar{\Phi}_\rho = \sigma^n_{k-1}(\rho)\bar{\Phi}_{\rho-n}, \qquad \bar{\Phi}_{\rho-n} \equiv 0 \quad \text{if } m_{k-1} < n, \tag{7.7}$$

where $\rho - n = (r_i - n)$, $1 \leqq i \leqq k-1$. $\rho - n$ is a partition only if $m_{k-1} > n$.

The eigenvalue $\sigma^n_{k-1}(\rho)$ may be computed via

$$\sigma^n_{k-1}(\rho) = \prod_{j=0}^{n-1} \sigma_{k-1}(\rho - j) \tag{7.8}$$

and is zero if $m_{k-1} < n$. Evaluating the right side of (7.4) by (7.5) and (7.7) gives

$$[D^n_k F](V) = \sigma^n_{k-1}(\rho) f_{m-n}(V)\bar{\Phi}_{\rho-n}(V). \tag{7.9}$$

Substituting (7.9) in (7.3), we obtain

$$D^n_k\{m, 0, \rho, h\} = \sigma^n_{k-1}(\rho) \int f_{m-n}(X'VX)\bar{\Phi}_{\rho-n}(X'VX)|X|^{2n} h(X)\, dX. \tag{7.10}$$

The first goal of the proof is accomplished in (7.10).

The second part of the proof is controlled by Lemma 2, since we need to calculate $L_\tau\{m, 0, \rho, h\}$ only at the matrix $V=0$. To do this, we use (7.10) to first evaluate $D_k^n\{m, 0, \rho, h\}$ at the matrix $V=\begin{pmatrix} V_{k-1} & 0 \\ 0 & 0 \end{pmatrix}$. Now

$$f_{m-n}\left(X'\begin{pmatrix} V_{k-1} & 0 \\ 0 & 0 \end{pmatrix}X\right)=f_{m-n}(X'_{k-1}V_{k-1}X_{k-1}) \tag{7.11a}$$

and

$$\bar{\Phi}_{\rho-n}\left(X'\begin{pmatrix} V_{k-1} & 0 \\ 0 & 0 \end{pmatrix}X\right)=\bar{\Phi}_{\rho[k-2]}(X'_{k-1}V_{k-1}X_{k-1})|X^*|^{2(m_{k-1}-n)}|V_{k-1}|^{m_{k-1}-n}, \tag{7.11b}$$

where X^* is defined in (7.2f). From (7.10) and (7.11a), (7.11b), we obtain

$$\begin{aligned}&[D_k^n\{m, 0, \rho, h_k\}]\left(V=\begin{pmatrix} V_{k-1} & 0 \\ 0 & 0 \end{pmatrix}\right)\\ &=\sigma_{k-1}^n(\rho)|V_{k-1}|^{m_{k-1}-n}\int f_{m-n}(X'_{k-1}V_{k-1}X_{k-1})\\ &\quad\cdot\bar{\Phi}_{\rho[k-2]}(X'_{k-1}V_{k-1}X_{k-1})h_{k-1}(X)\,dX\\ &=\sigma_{k-1}^n(\rho)\{m-n, m_{k-1}-n, \rho[k-2], h_{k-1}\}\end{aligned} \tag{7.12}$$

where

$$h_{k-1}(X)=|X^*|^{2(m_{k-1}-n)}|X|^{2n}h_k(X) \tag{7.13}$$

and $h_k(X)\equiv h(X)$. Note that the $\{\ \}$ function in (7.12) is now a function of the $(k-1)\times(k-1)$ matrix V_{k-1}. The second part of the proof is accomplished in (7.12).

In the third part of the proof, an expression is obtained for the effect of L_τ on the function $\{m, l, \rho, h\}$ at $V=0$. From Lemmas 1 and 2, with $\tau=(t_i)=[n_i]$, $1\leqq i\leqq k$,

$$\begin{aligned}[L_\tau\{m, l, \rho, h_k\}](V=0)&=b_k^l(\tau)[L_{\tau-l}\{m, 0, \rho, h_k\}](V=0)\\ &=b_k^l(\tau)[L_{\tau[k-1]}G](V_{k-1}=0)\end{aligned} \tag{7.14}$$

where

$$G(V_{k-1})=[D_k^{n_k-l}\{m, 0, \rho, h_k\}]\left(V=\begin{pmatrix} V_{k-1} & 0 \\ 0 & 0 \end{pmatrix}\right). \tag{7.15}$$

Evaluating (7.15) via (7.12), with $n=n_k-l$, allows (7.14) to be written as

$$\begin{aligned}[L_\tau\{m, l, \rho, h_k\}](V_k=0)&=b_k^l(\tau)\sigma_{k-1}^{n_k-l}(\rho)\\ &\quad\cdot[L_{\tau[k-1]}\{m-n_k+l, m_{k-1}-n_k+l, \rho[k-2], h_{k-1}\}](V_{k-1}=0).\end{aligned} \tag{7.16}$$

The third part of the proof is accomplished in (7.16).

The left side of (7.16) involves the $k\times k$ matrix V_k, and partitions τ and ρ. The right side of (7.16) involves, except for the constants $b_k^l(\tau)$ and $\sigma_{k-1}^{n_k-l}(\rho)$, the $(k-1)\times(k-1)$ matrix V_{k-1} and the truncated partitions $\tau[k-1]$ and $\rho[k-2]$. This enables us to proceed inductively into the fourth part of the proof. Indeed, repeating $k-i$ times the procedure which results in (7.16), we obtain:

$$[L_\tau\{m, l, \rho, h_k\}](V_k=0)=C_i[L_{\tau[i]}\{M_i, l_i, \rho[i-1], h_i\}](V_i=0) \tag{7.17}$$

where

$$M_i = m - \sum_{j=i+1}^{k} (j-i)n_j + \sum_{j=i+1}^{k-1} (j-i)m_j + (k-i)l, \tag{7.18}$$

$$l_i = \sum_{j=i}^{k-1} m_j - \sum_{j=i+1}^{k} n_j + l, \tag{7.19}$$

$$h_i(X) = h_k(X) \prod_{j=i+1}^{k} |X_{j-1}^*|^{2(m_{j-1}-n_j+l_j)} |X_j|^{2(n_j-l_j)}, \tag{7.20}$$

and

$$C_i = \prod_{j=i+1}^{k} b_j^l(\tau[j]) \sigma_{j-1}^{n_j-l_j}(\rho[j-1]). \tag{7.21}$$

In terms of the parts of the partitions τ and ρ, (7.18)-(7.21) are

$$M_i = m - \sum_{j=i+1}^{k} t_j + \sum_{j=i+1}^{k-1} r_i + (k-i)l, \tag{7.22}$$

$$l_i = r_i - t_{i+1} + l, \tag{7.23}$$

$$h_i(X) = h_k(X) \prod_{j=i+1}^{k} |X_{j-1}^*|^{2(r_{j-1}-t_j+l)} |X_j|^{2(t_j-r_j-l)}, \tag{7.24}$$

$$C_i = \prod_{j=i+1}^{k} b_j^{r_j-t_{j+1}+l}(\tau[j]) \sigma_{j-1}^{t_j-r_j-l}(\rho[j-1]). \tag{7.25}$$

When $i = k$, (7.17) is identically true, since $M_k = m$, $l_k = l$, $C_k = 1$. When $i = k-1$, (7.17) yields (7.16) since $M_{k-1} = m - n_k + l = m - t_k + l$, $l_{k-1} = m_{k-1} - n_k + l = r_{k-1} - t_k + l$, $C_{k-1} = b_k^l(\tau)\sigma_{k-1}^{n_k-l}(\rho) = b_k^l(\tau)\sigma_{k-1}^{t_k-l}(\rho)$.

Set $l = 0$ and $i = 0$ in (7.24) and (7.25). The exponents of $|X_{j-1}^*|$ and of $|X_j|$ must be nonnegative integers, yielding (7.1a). The conditions (7.1a) also ensure that M_i and l_i are nonnegative integers and that $C_i \neq 0$. The symbol $[L_{\tau[0]}\{M_0, l_0, \rho[-1], h_0\}] \equiv h_0$, where, from (7.24),

$$\begin{aligned} h_0(X) &= e^{-(1/2)\operatorname{tr} X'X} \prod_{j=1}^{k} |X_{j-1}^*|^{2(r_{j-1}-t_j)} |X_j|^{2(t_j-r_j)} \\ &= e^{-(1/2)\operatorname{tr} X'X} \prod_{j=1}^{k-1} |X_j^*|^{2(r_j-t_{j+1})} \prod_{j=1}^{k} |X_j|^{2(t_j-r_j)} \\ &= e^{-(1/2)\operatorname{tr} X'X} \Phi_\nu(X^*) \Phi_\mu(X), \end{aligned} \tag{7.26}$$

the integrand in (7.2a). The constant C_0 is given by

$$\begin{aligned} C_0 &= \prod_{j=1}^{k} b_j^{r_j-t_{j+1}}(\tau[j]) \sigma_{j-1}^{t_j-r_j}(\rho[j-1]) \\ &= \prod_{j=1}^{k-1} b_j^{r_j-t_{j+1}}(\tau[j]) \prod_{j=1}^{k-1} \sigma_j^{t_{j+1}-r_j+1}(\rho[j]) \end{aligned} \tag{7.27}$$

an expression which we now simplify.

The "σ-symbols" in (7.27) are evaluated as in (5.25)

$$\begin{aligned} \sigma_j(\rho[j]-i) &= \prod_{l=1}^{j} (j-l+2r_l-2r_{j+1}-2i) \\ &= \prod_{l=1}^{j} (R_l - R_{j+1} - 2i - 1). \end{aligned}$$

It follows that

$$\sigma_j^{t_{j+1}-r_{j+1}}(\rho[j]) = \prod_{i=0}^{t_{j+1}-r_{j+1}-1} \sigma_j(\rho[j]-i)$$

$$= \prod_{l=1}^{j} \prod_{i=0}^{t_{j+1}-r_{j+1}-1} (R_l - R_{j+1} - 2i - 1)$$

$$= 2^{j(t_{j+1}-r_{j+1})} \prod_{l=1}^{j} \left[\frac{1}{2}(R_l - R_{j+1} - 1), \frac{1}{2}(R_l - T_{j+1} + 1)\right]$$

and

$$\prod_{j=1}^{k-1} \sigma_j^{t_{j+1}-r_{j+1}}(\rho[j]) = 2^a \prod_{l \leqq j}^{k-1} \left[\frac{1}{2}(R_l - R_{j+1} - 1), \frac{1}{2}(R_l - T_{j+1} + 1)\right] \tag{7.28}$$

where $a = \sum_{j=1}^{k} j(t_{j+1} - r_{j+1})$.

The "b-symbols" in (7.27) are evaluated using (3.3) with $i = r_j - t_{j+1}$, $k = j$, and the partition $\tau[j]$ in place of τ. The lth part of $\tau[j]$ is $t_l - t_{j+1}$. To evaluate $b_j^{r_j - t_{j+1}}(\tau[j])$, the $T_l + k$ in (3.3) is replaced by $2(t_l - t_{j+1}) - l + j = T_l - T_{j+1} - 1$; the $T_l + k - 2i + 2$ in (3.3) is replaced by $T_l - T_{j+1} - 1 - 2r_j + 2t_{j+1} + 2 = T_l - R_j + 2$. We obtain

$$b_j^{r_j - t_{j+1}}(\tau[j]) = 2^{j(r_j - t_{j+1})} \prod_{l=1}^{j} \left[\frac{1}{2}(T_l - T_{j+1} - 1), \frac{1}{2}(T_l - R_j + 2)\right] \tag{7.29}$$

or

$$\prod_{j=1}^{k-1} b_j^{r_j - t_{j+1}}(\tau[j]) = 2^b \prod_{l \leqq j}^{k-1} \left[\frac{1}{2}(T_l - T_{j+1} - 1), \frac{1}{2}(T_l - R_j + 2)\right] \tag{7.30}$$

where $b = \sum_{j=1}^{k-1} j(r_j - t_{j+1})$. Note that $a + b = r$.

Multiplying (7.28) and (7.30) together, with $p = l(\rho)$, $k = p + 1$ (or any $k \geqq p + 1$) evaluates the C_0 in (7.27). Using this expression for C_0 and (7.26), the case $i = 0$, $l = 0$ of (7.17) is

$$[L_\tau\{m, 0, \rho, h\}](V_k = 0) = C_0 \int \Phi_\mu(X)\Phi_\nu(X^*)\, e^{-(1/2)\operatorname{tr} X'X}\, dX$$

which, with (3.2), yields (7.2a).

8. More eigenfunctions. Theorem 4 expresses the g-coefficients by integrals of the form

$$c_{\mu\nu} = (2\pi)^{-k^2/2} \int \Phi_\mu(X)\Phi_\nu(X^*)\, e^{-(1/2)\operatorname{tr} X'X}\, dX \tag{8.1}$$

where

$$\Phi_\mu(X) = \prod_{i=1}^{k} |X_i|^{e_i}, \qquad \mu = [e_1, e_2, \cdots, e_k]$$

and

$$\Phi_\nu(X^*) = \prod_{i=1}^{k-1} |X_i^*|^{f_i}, \quad \nu = [f_1, f_2, \cdots, f_{k-1}].$$

Here, X is a $k \times k$ matrix and X^* is the upper right $(k-1) \times (k-1)$ submatrix of X. When $f_i = 0$, $1 \leqq i \leqq k-1$, the integral (8.1) was evaluated in § 5. The procedure for the evaluation of (8.1) is similar, but more complicated. It uses two eigenfunction properties of the integrand in (8.1).

THEOREM 5. *Let* $f(X)=\Phi_\mu(X)\Phi_\nu(X^*)$. *Then* $f(X)$ *is an eigenfunction of all the bi-invariant operators.*

Proof. The function $\Phi_\mu(X)$ satisfies the functional equation (5.2). For any upper right triangular T_2, and T_1 of the form given in Lemma 5, the function $g(X)=\Phi_\nu(X^*)$ satisfies

$$g(T_2'XT_1)=w_1(T_1)w_2(T_2)g(X)$$

where w_i are functions of the matrices T_i, $1\leqq i\leqq 2$. Consequently, for two such upper right triangular matrices, $f(X)$ satisfies

$$f(T_2'XT_1)=w(T_1,T_2)f(X) \tag{8.2}$$

where w is a function of the matrices T_1 and T_2.

Let A be a matrix in the form given in Lemma 5. Since $a_i\neq 0$, $1\leqq i\leqq k$, $f(A)\neq 0$, because

$$\Phi_\mu(A)=a_1^{e_1}(a_1a_2)^{e_2}\cdots(a_1\cdots a_k)^{e_k}\neq 0$$

and

$$\Phi_\nu(A^*)=a_1^{f_1}(a_1a_2)^{f_2}\cdots(a_1\cdots a_{k-1})^{f_{k-1}}\neq 0.$$

Let $X=A$ in (8.2); we then obtain

$$f(T_2'AT_1)=w(T_1,T_2)f(A). \tag{8.3}$$

Using (8.3) to eliminate the w function from (8.2), we obtain for $f(X)$ the functional equation

$$f(T_2'XT_1)=f(T_2'AT_1)f(X)/f(A). \tag{8.4}$$

In (8.4), T_2 is any upper right triangular matrix; T_1 and A are matrices of the form given in Lemma 5. Operating on both sides of (8.4) with a bi-invariant operator L gives

$$[Lf](T_2'XT_1)=f(T_2'AT_1)[Lf](X)/f(A). \tag{8.5}$$

Setting $X=A$ in (8.5) yields

$$[Lf](T_2'AT_1)=\lambda f(T_2'AT_1)$$

where $\lambda=[Lf](A)/f(A)$. By Lemma 6, any matrix X, certain of whose subdeterminants do not vanish, can be represented in the form $T_2'AT_1$, where A is a constant matrix. Therefore, the eigenfunction property

$$[Lf](X)=\lambda f(X) \tag{8.6}$$

holds, except possibly at matrices X for which one or more of the subdeterminants vanish. As in Theorem 1, it follows that (8.6) also holds at any matrix X for which both sides of (8.6) are continuous, proving the theorem.

When $e_k=0$ and the first column vector of X is fixed, the function f defined in Theorem 5 may be viewed as a function of X^*. The second eigenfunction property relates to this function.

THEOREM 6. *Suppose that the first column vector of* X *is fixed and define*

$$F(X^*)=\prod_{i=1}^{k-1}|X_i|^{e_i}\prod_{i=1}^{k-1}|X_i^*|^{f_i}$$

viewed as a function of the variables in X^*, *the* $(k-1)\times(k-1)$ *upper right part of* X. *Then* F *is an eigenfunction of all the bi-invariant operators* (*acting on functions of* X^*).

Proof. We begin by proving that, if $x_{11} \neq 0$, there is a lower left invertible $(k-1)\times(k-1)$ triangular matrix S such that, if $Y^{\#}$ denotes the $(k-2)\times(k-2)$ lower left part of a $(k-1)\times(k-1)$ matrix Y, then

$$|X_{i+1}| = |[(SX^*)^{\#}]_i|, \qquad 1 \leqq i \leqq k-2. \tag{8.7}$$

In (8.7), the matrix S depends only on the entries, $\{x_{i1}\}$, $1 \leqq i \leqq k-1$ in the first column of X. Now Cauchy's composition law (Karlin (1968)), states that

$$|AB|_{\rho\kappa} = \sum_{\tau} |A|_{\rho\tau}|B|_{\tau\kappa}. \tag{8.8}$$

In (8.8), $\rho = (r_1, r_2, \cdots, r_i)$ and $\kappa = (k_1, k_2, \cdots, k_i)$ are two fixed partitions, each having i parts; τ runs over all partitions having i parts; $|C|_{\rho\kappa}$ is the $i \times i$ subdeterminant of the matrix C formed from rows $\{r_l\}$ and columns $\{k_l\}$; and A and B are any two square matrices of the same order. Define partitions ρ and $\langle j\rangle$, both of length i, by

$$\rho = (i+1, \cdots, 3, 2), \tag{8.9a}$$

$$\langle j\rangle = (i+1, \cdots, \hat{j}, \cdots, 2, 1). \tag{8.9b}$$

$\hat{j}$ indicates that j does not appear in the partition $\langle j\rangle$. If S is a lower left $(k-1)\times(k-1)$ triangular matrix, and $i \leqq k-2$, then

$$|S|_{\rho\tau} = 0, \quad \text{unless } \tau = \langle j\rangle, \qquad 1 \leqq j \leqq i+1. \tag{8.10}$$

Let $\kappa = (i, \cdots, 2, 1)$. Note that

$$|SX^*|_{\rho\kappa} = |[(SX^*)^{\#}]_i|. \tag{8.11}$$

In (8.8), set $A = S$ and $B = X^*$. Together with (8.10) and (8.11) we then obtain

$$|[(SX^*)^{\#}]_i| = \sum_{j=0}^{i+1} |S|_{\rho\langle j\rangle}|X^*|_{\langle j\rangle\kappa}. \tag{8.12}$$

If a $(k-1)\times(k-1)$ lower triangular matrix S can be found satisfying

$$|S|_{\rho\langle j\rangle} = (-)^{j+1}x_{j1}, \qquad 1 \leqq j \leqq i+1 \leqq k, \tag{8.13}$$

where ρ and $\langle j\rangle$—which depend on i—are given in (8.9a), (8.9b), then the right side of (8.12) is equal to $|X_{i+1}|$, which is the assertion (8.7). We now demonstrate the existence of such a matrix.

Let $C = \{a\}_\perp$, where $a = (x_{i1})$, $1 \leqq i \leqq k-1$. Since $a \neq 0$, C is a $(k-2)$-dimensional subspace of E_{k-1}, Euclidean $(k-1)$-dimensional space. From Lemma 4, with $l = k-2$, $n = k-1$, C has a trapezoidal basis: $C = \{a\}_\perp = \{s_j\}$, $2 \leqq j \leqq k-1$, where the final $k-1-j$ components of s_j are zero. Define the vector s_1 by $s_1 = (1, 0, 0, \cdots, 0) \in E_{k-1}$ and let S be the $(k-1)\times(k-1)$ matrix whose row vectors are s_j, $1 \leqq j \leqq k-1$. S is a triangular matrix. If $s_1 \notin C$, i.e., if $x_{11} \neq 0$, then S is invertible. For fixed $1 \leqq i \leqq k-1$ and any vector $y = (y_1, y_2, \cdots, y_{k-1}) \in E_{k-1}$, let $\bar{y} = (y_1, y_2, \cdots, y_i) \in E_i$. Then

$$\{\bar{a}_i\}_\perp = \{\bar{s}_j\}, \qquad 2 \leqq j \leqq i. \tag{8.14}$$

For, $\bar{a}_i \neq 0$, hence $\dim\{\bar{a}_i\}_\perp = i-1$, and certainly the $i-1$ independent vectors $\bar{s}_j$, $2 \leqq j \leqq i$, are orthogonal to $\bar{a}_i$. Equating the Grassmann coordinates of the subspaces appearing in (8.14) gives, for scalars $\lambda_i \neq 0$,

$$|S|_{\rho\langle j\rangle} = \lambda_i(-)^{j+1}x_{j1}, \qquad 1 \leqq j \leqq i+1 \leqq k,$$

and renormalizing the vectors s_i results in a invertible triangular matrix S satisfying (8.13). This concludes the proof of (8.7).

We now use (8.7) to obtain the assertion of the theorem. Let $Y = X^*$. Then (8.7) together with $|(SY)_i| = |S_i||Y_i|$, $1 \leqq i \leqq k-1$ implies that the polynomial F defined in Theorem 6 can be written as

$$F(Y) = x_{11}^{e_1} w(S) H(SY) \tag{8.15}$$

where

$$H(Y) = \prod_{i=2}^{k-1} |Y_i^{\#}|^{e_i} \prod_{i=1}^{k-1} |Y_i|^{f_i} \tag{8.16}$$

and

$$w(S) = \prod_{i=1}^{k-1} |S_i|^{-f_i}.$$

But the polynomial $H(Y)$ is the analogue of the polynomial $f(X)$ in Theorem 5. Therefore, H is an eigenfunction of all the bi-invariant operators (acting on functions of $Y = X^*$) and also so is the left-translate of H and, by (8.15), the function $F(Y) = F(X^*)$.

9. More eigenvalues. In this section, we compute the eigenvalues of the functions $f(X)$ and $F(X^*)$ defined in the previous section.

The sum of two partitions is defined by component-wise addition. (This definition holds for both representations of partitions (§ 1).) Thus α, the sum of the two partitions μ and ν defined in the previous section, is

$$\alpha = \mu + \nu = [e_1 + f_1, e_2 + f_2, \cdots, e_{k-1} + f_{k-1}, e_k], \qquad k \geqq 2. \tag{9.1}$$

If $k = 1$, ν is defined to be a "null partition," i.e., a partition without any parts. In this case, $\alpha = \mu = [e_1]$. If $k \geqq 3$, define also the partition μ^* by

$$\mu^* = [e_2, \cdots, e_{k-1}]. \tag{9.2}$$

β, the sum of μ^* and ν, is

$$\beta = \mu^* + \nu = [e_2 + f_1, e_3 + f_2, \cdots, e_{k-1} + f_{k-2}, f_{k-1}], \qquad k \geqq 3. \tag{9.3}$$

If $k \leqq 2$, μ^* is the null partition. In this case, if $k = 2$, $\beta = \nu = [f_1]$; if $k = 1$, β itself is the null partition.

Using the above notation and that of the previous section, the next theorem gives the eigenvalues of the functions $f(X)$ and $F(X^*)$, defined as in Theorems 5 and 6, respectively.

THEOREM 7.

$$|X| D_X f = \sigma_k(\tfrac{1}{2}\alpha) f, \qquad \alpha = \mu + \nu. \tag{9.4}$$ (a)

$$|X^*| D_{X^*} F = \sigma_{k-1}(\tfrac{1}{2}\beta) F, \qquad \beta = \mu^* + \nu. \tag{9.5}$$ (b)

Proof. (a) In (8.4), let $T_1 = I$. Also, call A_1 the matrix obtained by setting $a_i = 1$, $1 \leqq i \leqq k$, in the matrix A. We obtain a functional equation for f

$$f(SX) = f(SA_1) f(X), \quad S = (s_{ij}) \text{ lower triangular.} \tag{9.6}$$

But

$$\begin{aligned} \Phi_\mu(SA_1) &= s_{11}^{e_1} (s_{11} s_{22})^{e_2} \cdots (s_{11} s_{22} \cdots s_{kk})^{e_k} \\ &= \Phi_{\mu/2}(SS') \end{aligned}$$

and

$$\Phi_\nu((SA_1)^*) = s_{11}^{f_1}(s_{11}s_{22})^{f_2}\cdots(s_{11}s_{22}\cdots s_{k-1,k-1})^{f_{k-1}}$$
$$= \Phi_{\nu/2}(SS')$$

implying $f(SA_1) = \Phi_\mu(SA_1)\Phi_\nu((SA_1)^*) = \Phi_{(\mu+\nu)/2}(SS')$. Therefore (9.6) can be written as

$$f(SX) = \Phi_{(\mu+\nu)/2}(SS')f(X), \quad S \text{ lower left triangular}, \tag{9.7}$$

which is condition (5.4b) of Lemma 7. By Theorem 5, $f(X)$ is an eigenfunction of $|X|D_X$. Now assume that the e_i and f_i are even integers, in which case $f(X) \geqq 0$. The assertion of part (a) then follows, in this case, from Lemma 7. To prove (9.4) in the general case, when the e_i and f_i are any integers, one can argue as in the proof of Theorem 2.

(b) Let $X^{**} = (x_{ij})$, $2 \leqq i, j \leqq k$, a submatrix of X^*. If $Y = X^*$, then $Y^{\#} = X^{**}$. It was shown in the proof of Theorem 6 that $F(X^*)$ is a multiple of a left-translate of a function H defined in (8.16), which in the present notation can be written as

$$H(X^*) = \Phi_\nu(X^*)\Phi_{\mu^*}(X^{**}). \tag{9.8}$$

The eigenvalue of $F(X^*)$ is the same as the eigenvalue of $H(X^*)$. Since $H(X^*)$ is the analogue of the function $f(X)$, part (a) of this theorem can be applied to obtain part (b).

10. Evaluation of integrals and g-coefficients. In this section, after evaluating integrals of the form (10.1a) below, we prove a simple formula for the g-coefficients $g^{\tau}_{(m)\rho}$.

THEOREM 8. *Let $\mu = [e_1, e_2, \cdots, e_k]$ and $\nu = [f_1, f_2, \cdots, f_{k-1}]$ be two partitions whose parts are even. The integral*

$$c_{\mu\nu} = (2\pi)^{-k^2/2}\int \Phi_\mu(X)\Phi_\nu(X^*)\, e^{-(1/2)\operatorname{tr} X'X}\, dX \tag{10.1a}$$

is given by

$$c_{\mu\nu} = 2^t \prod_{l\leqq j}^{q}\left[\frac{1}{2}(T_l - R_j - 1), \frac{1}{2}(T_l - T_j + 1)\right]$$
$$\cdot\left[\frac{1}{2}(R_l - T_{j+1} - 2), \frac{1}{2}(R_l - R_j + 1)\right], \qquad q = l(\mu), \tag{10.1b}$$

where

$$2\tau = \mu + \nu, \quad \tau = (t_j), \quad t = \sum_{j=1}^{q} t_j, \tag{10.2a}$$

$$2\rho = \mu^* + \nu, \qquad \rho = (r_j), \tag{10.2b}$$

$T_j = 2t_j - j$, $R_j = 2r_j - j$, *and* $\mu^* = [e_2, e_3, \cdots, e_{q-1}]$.

Proof. The function $g(X) = e^{-(1/2)\operatorname{tr} X'X}$ satisfies $D_X g = (-)^k|X|g$. Using (5.13) we express $c_{\mu\nu} = c_{\mu\nu}(k)$ by

$$c_{\mu\nu}(k) = (2\pi)^{-k^2/2}(-)^k\int \Phi_{\mu-1}(X)\Phi_\nu(X^*)D_X g\, dX$$
$$= (2\pi)^{-k^2/2}\int D_X[\Phi_{\mu-1}(X)\Phi_\nu(X^*)]g(X)\, dX \tag{10.3}$$
$$= (2\pi)^{-k^2/2}\int |X|^{-1}|X|D_X[\Phi_{\mu-1}(X)\Phi_\nu(X^*)]g(X)\, dX.$$

From Theorem 7, part (a), and (10.3), we obtain

$$c_{\mu\nu}(k) = (2\pi)^{-k^2/2}\sigma_k\left(\frac{\mu+\nu-1}{2}\right)\int \Phi_{\mu-2}(X)\Phi_\nu(X^*)g(X)\,dX.$$

Repeating this process $e_k/2$ times, we get

$$c_{\mu\nu}(k) = (2\pi)^{-k^2/2}\prod_{i=1}^{e_k/2}\sigma_k\left(\frac{\mu+\nu-2i+1}{2}\right)\int \Phi_{\mu[k-1]}(X)\Phi_\nu(X^*)g(X)\,dX. \tag{10.4}$$

The variables in the kth row of X do not appear in $\Phi_{\mu[k-1]}(X)\Phi_\nu(X^*)$. Fix the variables in the first column of X and integrate with respect to the remaining variables (those of X^*). By (5.13), with X^* in place of X, and Theorem 7, part (b), we further reduce the integral in (10.4), obtaining

$$c_{\mu\nu}(k) = (2\pi)^{-k^2/2}\left[\prod_{i=1}^{e_k/2}\sigma_k\left(\frac{\mu+\nu-2i+1}{2}\right)\right]\sigma_{k-1}\left(\frac{\mu^*+\nu-1}{2}\right)$$
$$\cdot\int \Phi_{\mu[k-1]}(X)\Phi_{\nu-2}(X^*)g(X)\,dX.$$

Again repeating the process $f_{k-1}/2$ times, we obtain

$$c_{\mu\nu}(k) = (2\pi)^{-k^2/2}\prod_{i=1}^{e_k/2}\sigma_k\left(\frac{\mu+\nu-2i+1}{2}\right)\prod_{i=1}^{f_{k-1}/2}\sigma_{k-1}\left(\frac{\mu^*+\nu-2i+1}{2}\right)$$
$$\cdot\int \Phi_{\mu[k-1]}(X)\Phi_{\nu[k-2]}(X^*)g(X)\,dX. \tag{10.5}$$

The variables in the kth row and column of X do not appear in $\Phi_{\mu[k-1]}(X)\Phi_{\nu[k-2]}(X^*)$. Integrating these $2k-1$ variables in integral (10.5) yields a recursion relation for $c_{\mu\nu}$

$$c_{\mu\nu}(k) = \prod_{i=1}^{e_k/2}\sigma_k\left(\frac{\mu+\nu-2i+1}{2}\right)$$
$$\cdot\prod_{i=1}^{f_{k-1}/2}\sigma_{k-1}\left(\frac{\mu^*+\nu-2i+1}{2}\right)c_{\mu[k-1],\nu[k-1]}(k-1). \tag{10.6}$$

Note that in (10.6), μ, ν, and μ^* are partitions of at most k, $k-1$, and $k-2$ parts, respectively.

The recursion relation (10.6) immediately yields a product formula for $c_{\mu\nu}$

$$c_{\mu\nu} = \prod_{j=1}^{q}\prod_{i=1}^{e_j/2}\sigma_j\left(\frac{\mu[j]+\nu[j-1]-2i+1}{2}\right)$$
$$\cdot\prod_{j=1}^{q-1}\prod_{i=1}^{f_j/2}\sigma_j\left(\frac{\mu^*[j-1]+\nu[j]-2i+1}{2}\right). \tag{10.7}$$

In (10.7), $\mu^*[j-1]=[e_2, e_3, \cdots, e_j]$, a partition of at most $j-1$ parts. (As usual, all the empty products, those with $e_j=0$, and $f_j=0$, do not appear in (10.7).)

We now simplify (10.7) by writing μ and ν as

$$\mu = 2[t_j - r_j], \tag{10.8a}$$

$$\nu = 2[r_j - t_{j+1}] \tag{10.8b}$$

which are the inverse forms of (10.2a), (10.2b). From (10.8a) we have $\mu[j]+\nu[j-1]=2[t_1-t_2, t_2-t_3, \cdots, t_{j-1}-t_j, t_j-r_j]$. By the same technique already used to evaluate σ-symbols

$$\prod_{j=1}^{q}\prod_{i=1}^{e_j/2}\sigma_j((\mu[j]+\nu[j-1]-2i+1)/2)=\prod_{j=1}^{q}\prod_{l=1}^{j}\prod_{i=1}^{t_j-r_j}(j-l+2(t_l-r_j)-2i+1)$$
$$=2^c\prod_{l\leqq j}^{q}\left[\frac{1}{2}(T_l-R_j-1),\frac{1}{2}(T_l-T_j+1)\right] \tag{10.9}$$

where $c=\sum_{j=1}^{q}j(t_j-r_j)$. Again, from (10.8b) we have $\mu^*[j-1]+\nu[j]=2[r_1-r_2, r_2-r_3, \cdots, r_{j-1}-r_j, r_j-t_{j+1}]$. Therefore,

$$\prod_{j=1}^{q}\prod_{i=1}^{f_j/2}\sigma_j((\mu^*[j-1]+\nu[j]-2i+1)/2)$$
$$=\prod_{j=1}^{q}\prod_{l=1}^{j}\prod_{i=1}^{r_j-t_{j+1}}(j-l+2(r_l-t_{j+1})-2i+1) \tag{10.10}$$
$$=2^b\prod_{l\leqq j}^{q}\left[\frac{1}{2}(R_l-T_{j+1}-2),\frac{1}{2}(R_l-R_j+1)\right]$$

where $b=\sum_{j=1}^{q}j(r_j-t_{j+1})$. Note that $c+b=t$. From (10.7), after multiplying (10.9) and (10.10) together, we get the assertion of the theorem.

THEOREM 9. *Suppose that $\tau=(t_i)$, and $\rho=(r_i)$, are partitions of t and r, respectively. If*

$$t_i\geqq r_i\geqq t_{i+1},\qquad 1\leqq i\leqq p=l(\rho) \tag{10.11a}$$

and

$$t=m+r, \tag{10.11b}$$

then the g-coefficient $g^{\tau}_{(m)\rho}$ defined in (1.2) *is given by*

$$\binom{t}{m}g^{\tau}_{(m)\rho}=\frac{m!}{(2m)!}\prod_{l=1}^{p+1}\frac{[2\delta_l-1,1]_2}{\delta_l!}\prod_{l<j}^{p+1}(R_l-R_j)$$
$$\cdot\prod_{l<j}^{p+2}[T_l-R_j-1, R_l-T_j+1]_2\Big/\prod_{l<j}^{p+1}[T_l-R_j, R_l-T_j]_2 \tag{10.12}$$

where $\delta_l=t_l-r_l$, $R_l=2r_l-l$, $T_l=2t_l-l$. The factorial symbol $[b, c]_2$ is defined by

$$[b, c]_2=b(b-2)(b-4)\cdots c \quad \text{if } b-c=2n\geqq 0,$$
$$[b, c]_2=1 \quad \text{if } b-c=-2.$$

If τ and ρ do not satisfy (10.11a), (10.11b), *then $g^{\tau}_{(m)\rho}=0$. If $g^{\tau}_{(m)\rho}\neq 0$, then*

$$l(\rho)\leqq l(\tau)\leqq l(\rho)+1. \tag{10.13}$$

Proof. From Theorem 4 and Theorem 8 (with $q=p+1$)

$$2^t t!\,g^{\tau}_{(m)\rho}=m!\,a_{(m)}a_{\rho}2^{2t}(2\pi)^{k^2/2}$$
$$\cdot\prod_{l\leqq j}^{p}\frac{\Gamma(\frac{1}{2}(R_l-R_{j+1}+1))}{\Gamma(\frac{1}{2}(R_l-T_{j+1}+1))}\frac{\Gamma(\frac{1}{2}(T_l-T_{j+1}+1))}{\Gamma(\frac{1}{2}(T_l-R_j+2))}$$
$$\cdot\prod_{l\leqq j}^{p+1}\frac{\Gamma(\frac{1}{2}(T_l-R_j+1))}{\Gamma(\frac{1}{2}(T_l-T_j+1))}\frac{\Gamma(\frac{1}{2}(R_l-T_{j+1}))}{\Gamma(\frac{1}{2}(R_l-R_j+1))}$$

which simplifies to

$$
\begin{aligned}
t!\,g^{\tau}_{(m)\rho} &= m!\,a_{(m)}a_{\rho}2^{t}(2\pi)^{(p+1)^2/2}\pi^{-(p+1)(p+2)/2} \\
&\quad\cdot\prod_{l=1}^{p+1}\Gamma\left(\frac{1}{2}(T_l-R_l+1)\right)\Gamma\left(\frac{1}{2}(R_l+p+2)\right) \\
&\quad\cdot\prod_{l\leqq j}^{p}\frac{\Gamma(\frac{1}{2}(T_l-R_{j+1}+1))}{\Gamma(\frac{1}{2}(R_l-T_{j+1}+1))}\,\frac{\Gamma(\frac{1}{2}(R_l-T_{j+1}))}{\Gamma(\frac{1}{2}(T_l-R_j+2))} \\
&= m!\,a_{(m)}a_{\rho}2^{t}(2\pi)^{(p+1)^2/2}\pi^{-(p+1)(p+2)/2} \\
&\quad\cdot\prod_{l=1}^{p+1}\Gamma\left(\frac{1}{2}(T_l-R_l+1)\right)\Gamma\left(\frac{1}{2}(R_l+p+2)\right) \\
&\quad\cdot\prod_{l<j}^{p+1}\frac{\Gamma(\frac{1}{2}(T_l-R_j+1))}{\Gamma(\frac{1}{2}(R_l-T_j+1))}\,\frac{\Gamma(\frac{1}{2}(R_l-T_j))}{\Gamma(\frac{1}{2}(T_l-R_j+2))} \\
&\quad\cdot\prod_{l\leqq j}^{p}\frac{\Gamma(\frac{1}{2}(T_l-R_{j+1}+2))}{\Gamma(\frac{1}{2}(T_l-R_j+2))} \qquad (10.14) \\
&= m!\,a_{(m)}a_{\rho}2^{t}(2\pi)^{(p+1)^2/2}\pi^{-(p+1)(p+2)/2} \\
&\quad\cdot\prod_{l=1}^{p+1}\frac{\Gamma(\frac{1}{2}(T_l-R_l+1))}{\Gamma(\frac{1}{2}(T_l-R_l+2))}\Gamma\left(\frac{1}{2}(T_l+p+3)\right)\Gamma\left(\frac{1}{2}(R_l+p+2)\right) \\
&\quad\cdot\prod_{l<j}^{p+1}\frac{\Gamma(\frac{1}{2}(T_l-R_j+1))}{\Gamma(\frac{1}{2}(R_l-T_j+1))}\,\frac{\Gamma(\frac{1}{2}(R_l-T_j))}{\Gamma(\frac{1}{2}(T_l-R_j+2))} \\
&= m!\,a_{(m)}a_{\rho}2^{t}(2\pi)^{(p+1)^2/2}\pi^{-(p+1)(p+2)/2} \\
&\quad\cdot\prod_{l=1}^{p+1}\frac{\Gamma(\frac{1}{2}(T_l-R_l+1))}{\Gamma(\frac{1}{2}(T_l-R_l+2))}\Gamma\left(\frac{1}{2}(R_l+p+3)\right)\Gamma\left(\frac{1}{2}(R_l+p+2)\right) \\
&\quad\cdot\prod_{l<j}^{p+2}\frac{\Gamma(\frac{1}{2}(T_l-R_j+1))}{\Gamma(\frac{1}{2}(R_l-T_j+1))}\prod_{l<j}^{p+1}\frac{\Gamma(\frac{1}{2}(R_l-T_j))}{\Gamma(\frac{1}{2}(T_l-R_j+2))}.
\end{aligned}
$$

When $m=0$, $\tau=\rho$, $t=r$ and $g^{\rho}_{(0)\rho}=1$. In this case, (10.14) is

$$
\begin{aligned}
r! &= a_{(0)}a_{\rho}2^{r}(2\pi)^{(p+1)^2/2}\pi^{-(p+1)(p+2)/2}\pi^{(p+1)(p+2)/4} \\
&\quad\cdot\prod_{l=1}^{p+1}\Gamma\left(\frac{1}{2}(R_l+p+3)\right)\Gamma\left(\frac{1}{2}(R_l+p+2)\right) \qquad (10.15) \\
&\quad\cdot\prod_{l<j}^{p+1}(2/(R_l-R_j)).
\end{aligned}
$$

Dividing (10.14) by (10.15) eliminates a_{ρ} and other terms from (10.13), giving

$$
\begin{aligned}
a_{(0)}t!\,g^{\tau}_{(m)\rho} &= r!\,m!\,a_{(m)}2^{m}\pi^{-(p+1)(p+2)/4}\prod_{l=1}^{p+1}\frac{\Gamma(\frac{1}{2}(T_l-R_l+1))}{\Gamma(\frac{1}{2}(T_l-R_l+2))} \\
&\quad\cdot\prod_{l<j}^{p+2}\frac{\Gamma(\frac{1}{2}(T_l-R_j+1))}{\Gamma(\frac{1}{2}(R_l-T_j+1))}\prod_{l<j}^{p+1}\frac{(R_l-R_j)\Gamma(\frac{1}{2}(R_l-T_j))}{2\Gamma(\frac{1}{2}(T_l-R_j+2))}. \qquad (10.16)
\end{aligned}
$$

Now if $b-c=2n\geqq -2$, then $[b,c]_2=2^{(b-c+2)/2}\Gamma(1+b/2)/\Gamma(c/2)$ $(c\neq 0,-2,-4,\cdots)$. Also, $\Gamma(\frac{1}{2}(T_l-R_l+1))=\Gamma(\frac{1}{2})2^{r_l-t_l}[2\delta_l-1,1]_2$ and $\Gamma(\frac{1}{2}(T_l-R_l+2))=\delta_l!$ Hence, (10.16) can be written as

$$
\begin{aligned}
a_{(0)}t!\,g^{\tau}_{(m)\rho}=r!\,m!\,a_{(m)}\prod_{l=1}^{p+1}\frac{[2\delta_l-1,1]_2}{\delta_l!}\\
\cdot\frac{\prod_{l<j}^{p+2}[T_l-R_j-1,R_l-T_j+1]_2}{\prod_{l<j}^{p+1}(R_l-R_j)[T_l-R_j,R_j-T_j]_2}.
\end{aligned}
\tag{10.17}
$$

Finally, $a_{(m)}$ is given in § 2 (with $q_1=1$ and $k=p+1$) by

$$a_{(m)}=C_{(m)}(I_{p+1})/((2\pi)^{(p+1)/2}2^m((p+1)/2)_m)=2^m m!/((2m)!(2\pi)^{k/2}),$$

so

$$a_{(m)}/a_{(0)}=2^m m!/(2m)!. \tag{10.18}$$

Substituting (10.18) in (10.17) gives the formula (10.12) of the theorem.

The interlacing condition (10.11a) is from Theorem 4. The inequality (10.13) is the case $l(m)=1$ of (6.15).

As an example of the theorem, let $m=2$, $\rho=(3,1,1)=(r_1,r_2,r_3)$, with $r=5$ and $p=3$. With $r_4=0$, the only partitions of $7=t=m+r$ satisfying (10.11a),

$$t_1\geqq 3\geqq t_2\geqq 1\geqq t_3\geqq 1\geqq t_4\geqq 0\geqq t_5,\qquad t_1+t_2+t_3+t_4=7,$$

are (5, 1, 1), (4, 2, 1), (4, 1, 1, 1), (3, 3, 1), and (3, 2, 1, 1).

The above five partitions τ are the only ones for which the coefficient $g^{\tau}_{(2)\rho}\neq 0$, as is also seen from Khatri and Pillai's (1968) table. Let $\tau=(4,2,1)$. Then from

	1	2	3	4	5
R_i	5	0	−1	−4	−5
T_i	7	2	−1	−4	−5
δ_i	1	1	0	0	0

we calculate

$$\prod_{l<j}^{4}(R_l-R_j)=5\cdot 6\cdot 9\cdot 1\cdot 4\cdot 3=a,$$

$$\prod_{l=1}^{4}[2\delta_l-1,1]_2=[1,1]_2\cdot[1,1]_2\cdot[-1,1]_2\cdot[-1,1]_2=1=b,$$

$$\prod_{l=1}^{4}\delta_l!=1\cdot 1\cdot 1\cdot 1=1=c,$$

	1,2	1,3	1,4	1,5	2,3	2,4	2,5	3,4	3,5	4,5
T_l-R_j	7	8	11	12	3	6	7	3	4	1
R_l-T_j	3	6	9	10	1	4	5	3	4	1

$$\prod_{l<j}^{5}[T_l-R_j-1, R_l-T_j+1]_2$$

$$=[6,4]_2[7,7]_2[10,10]_2[11,11]_2[2,2]_2[5,5]_2[6,6]_2[2,4]_2[3,5]_2[0,2]_2$$

$$=6\cdot 4\cdot 7\cdot 10\cdot 11\cdot 2\cdot 5\cdot 6=d,$$

$$\prod_{l<j}^{4}[T_l-R_j, R_l-T_j]=[7,3]_2[8,6]_2[11,9]_2[3,1]_2[6,4]_2[3,3]_2$$

$$=7\cdot 5\cdot 3\cdot 8\cdot 6\cdot 11\cdot 9\cdot 3\cdot 6\cdot 4\cdot 3=e,$$

$$m!\Big/(2m)!\binom{t}{m}=2!\Big/4!\binom{7}{2}=\frac{2\cdot 2}{4\cdot 3\cdot 2\cdot 7\cdot 6}=f.$$

According to (10.12),

$$g^{\tau}_{(m)\rho}=\frac{f\times b\times a\times d}{c\times e}$$

$$=\frac{2\cdot 2\cdot 5\cdot 6\cdot 9\cdot 4\cdot 3\cdot 6\cdot 4\cdot 7\cdot 10\cdot 11\cdot 2\cdot 5\cdot 6}{4\cdot 3\cdot 2\cdot 7\cdot 6\cdot 7\cdot 5\cdot 3\cdot 8\cdot 6\cdot 11\cdot\ 9\cdot 3\cdot 6\cdot 4}$$

$$=\frac{25}{189},$$

as given in Khatri and Pillai's (1968) table.

11. Special cases and applications.

(a) The linearization of tr $VC_\rho(V)$.

This is the case $m=1$ of Theorem 9, solved in Kushner (1985). Here (10.11a), (10.11b) imply that $g^{\tau}_{(1)\rho}\neq 0$ if and only if, for some $1\leqq i\leqq l(\rho)+1$, $\tau=\rho_i=(r_1, r_2, \cdots, r_{i-1}, r_i+1, r_{i+1}, \cdots)$ or $t_l=r_l+\delta^i_l$ (Kronecker delta). It follows that $T_l=R_l+2\delta^i_l$ and $\delta_l=\delta^i_l$. Also,

$$[T_l-R_j, R_l-T_j]_2=[R_l-R_j+2\delta^i_l, R_l-R_j-2\delta^i_l]_2$$

and

$$[T_l-R_j-1, R_l-T_j+1]_2=[R_l-R_j-1+2\delta^l_i, R_l-R_j-2\delta^l_i]_2.$$

Evaluating the product (10.12) leads to

$$(1+r)g^{\rho_i}_{(1)\rho}=\prod_{j=1}^{p+2}(R_i-R_j+1)\Big/\prod_{j=1}^{p+1}(R_i-R_j+2) \tag{11.1}$$

where $p=l(\rho)$ and ρ is a partition of r. Equation (11.1) is an equivalent, though more compact, form of the formula given in Theorem 1 of Kushner (1985).

(b) The linearization of the product of polynomials in one variable. Variations of the method of this paper—but by no means the full machinery—result in formulas for the linearization coefficients for some of the problems described in Askey (1975).

(c) An integral. A special case of the integral

$$c_{2\tau,\nu}=(2\pi)^{-k^2/2}\int \Phi_{2\tau}(X)|\det X|^{\nu}\, e^{-(1/2)\operatorname{tr} X'X}\, dX, \qquad \nu=n-k-1 \tag{11.2}$$

occurs in Schwager and Margolin (1982, p. 950). In (11.2), $|\det X|$ denotes the absolute value of $\det X$. When $\nu = 2e$ is an even integer, the integral (11.2) is given by Theorem 3 since

$$\Phi_{2\tau}(X)|\det X|^{\nu} = \Phi_{2(\tau+e)}(X),$$

implies

$$c_{2\tau,\nu} = c_{2(\tau+e)}, \qquad \nu = 2e. \tag{11.3}$$

The partition $(l_i) = \lambda = \tau + e$, having k parts, is given by

$$(l_i) = \lambda = \tau + e = (t_1 + e, t_2 + e, \cdots, t_p + e, e, e, \cdots, e)$$

from which we obtain

$$L_i = T_i + 2e, \qquad 1 \leqq i \leqq k,$$

$$L_{k+1} = T_{k+1} = -(k+1)$$

where $L_i = 2l_i - i$, $T_i = 2t_i - i$ (and $t_i = 0$, $p < i \leqq k+1$). Hence, $L_l - L_i = T_l - T_i$ if $1 \leqq i \leqq k$. Theorem 3 gives

$$\begin{aligned} c_{2(\tau+e)} = c_{2\lambda} &= \prod_{l \leqq i}^{k-1} [T_l - T_{i+1} - 2, T_l - T_i + 1]_2 \prod_{l=1}^{k} [T_l + 2e - T_{k+1} - 2, T_l - T_k + 1]_2 \\ &= \prod_{l \leqq i}^{k} [T_l - T_{i+1} - 2, T_l - T_i + 1]_2 \prod_{l=1}^{k} [T_l + 2e - T_{k+1} - 2, T_l - T_{k+1}]_2 \\ &= \prod_{l \leqq i}^{p} [T_l - T_{i+1} - 2, T_l - T_i + 1]_2 \prod_{l=1}^{k} [T_l + \nu + k - 1, T_l + k + 1]_2 \\ &= c_{2\tau} \prod_{l=1}^{k} [T_l + \nu + k - 1, T_l + k + 1]_2. \end{aligned} \tag{11.4}$$

In general, $c_{2\tau,\nu}$ can be computed using the substitution $X = HT$, as in Theorem 3. We obtain

$$c_{2\tau,\nu} = (2\pi)^{-k^2/2} 2^k g_k \int \Phi_{2\tau}(T)|T|^{\nu} e^{-(1/2)\operatorname{tr} T'T} \prod_{i=1}^{k} t_{ii}^{k-i}\, dT \int \Phi_{2\tau}(H)\, dH. \tag{11.5}$$

In (11.5), $|T|$ is, as usual, the determinant of T. In the same way that (5.18) followed from (5.17), we write the T-integral in (11.5) as an integral over the $k \times k$ positive definite matrices, and (11.5) becomes

$$\begin{aligned} c_{2\tau,\nu} &= (2\pi)^{-k^2/2} g_k \int \Phi_{\tau}(V)|V|^{(n-k-2)/2} e^{-(1/2)\operatorname{tr} V}\, dV \int \Phi_{2\tau}(H)\, dH \\ &= 2^{k(n-1-k)/2} \left[\prod_{i=1}^{k} \Gamma((n-k-1+k-i+1)/2)/\Gamma(i/2) \right] \\ &\quad \cdot [E_{n-1}\Phi_{\tau}](I_k) \int \Phi_{2\tau}(H)\, dH \end{aligned} \tag{11.6}$$

$$= 2^{k\nu/2} \left[\prod_{i=1}^{k} \Gamma((\nu+i)/2)\Gamma(i/2) \right] 2^t((n-1)/2)_{\tau} [c_{2\tau}/2^t(k/2)_{\tau}] \tag{11.7}$$

$$= 2^{k\nu/2} c_{2\tau} \left[\prod_{i=1}^{k} \Gamma((\nu+i)/2)/\Gamma(i/2) \right] ((n-1)/2)_{\tau}/(k/2)_{\tau}. \tag{11.8}$$

In (11.6), E_{n-1} is the expectation operator in the Wishart distribution, $W_k(n-1, \Sigma)$. Equation (11.7) follows from (11.8) from (D.2) of § 1 and Theorem 3. For any integral ν, $c_{2\tau,\nu}$ is given by (11.8). If ν is an even integer, after simplification, (11.8) reduces to (11.4). The case considered by Schwager and Margolin (1982) was $\tau = (4)$.

Acknowledgments. I would like to thank Donald Richards for suggesting application (c) and the referee for suggesting a shortening of the proof of (5.21).

REFERENCES

R. ASKEY (1975), *Orthogonal Polynomials and Special Functions*, CBMS-NSF Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, PA.

A. G. CONSTANTINE (1963), *Some noncentral distribution problems in multivariate analysis*, Ann. Math. Statist., 34, pp. 1270–1285.

——— (1966), *The distribution of Hotelling's generalized* T_0^2, Ann. Math. Statist., 37, pp. 215–225.

R. FARRELL (1976), *Techniques of Multivariate Calculation*, Lecture Notes in Mathematics 520, Springer-Verlag, New York, Berlin.

T. HAYAKAWA (1967), *On the distribution of the maximum latent root of a positive definite symmetric random matrix*, Ann. Inst. Statist. Math., 19, pp. 1–17.

E. HYLLERAAS (1962), *Linearization of products of Jacobi polynomials*, Math. Scand., 10, pp. 189–200.

A. T. JAMES (1960), *The distribution of the latent roots of the covariance matrix*, Ann. Math. Statist., 31, pp. 151–158.

——— (1961), *Zonal polynomials of the real positive definite symmetric matrices*, Ann. of Math., 74, pp. 456–469.

——— (1964), *Distribution of matrix variates and latent roots derived from normal samples*, Ann. Math. Statist., 35, pp. 475–501.

S. KARLIN (1968), *Total Positivity*, Stanford University Press, Stanford, CA.

L. KATES (1981), *Zonal polynomials*, Ph.D. thesis, Princeton University, Princeton, NJ.

C. G. KHATRI AND K. C. S. PILLAI (1968), *On the noncentral distributions of two test criteria in multivariate analysis of variance*, Ann. Math. Statist., 39, pp. 215–226.

D. KIKUCHI (1981), *Comparison of the James and Farrell approaches to zonal polynomials*, Technical Report 258, Department of Statistics, Ohio State University, Columbia, OH.

H. B. KUSHNER (1980), *Wishart expectation operators and invariant differential operators*, Ph.D. thesis, Yeshiva University, New York, NY.

H. B. KUSHNER, A. LEBOW, AND M. MEISNER (1981), *Eigenfunctions of expected value operators in the Wishart distribution*, II, J. Multivariate Anal., 11, pp. 418–433.

H. B. KUSHNER AND M. MEISNER (1984), *Formulas for zonal polynomials*, J. Multivariate Anal., 14, pp. 336–347.

H. B. KUSHNER (1985), *On the expansion of* $C_\rho^*(V+I)$ *as a sum of zonal polynomials*, J. Multivariate Anal., 17, pp. 84–98.

S. LANG (1966), *Linear Algebra*, Addison-Wesley, Reading, MA.

H. MAASS (1971), *Siegel's Modular Forms and Dirichlet Series*, Lecture Notes in Mathematics 216, Springer-Verlag, New York, Berlin.

I. G. MACDONALD (1979), *Symmetric Functions and Hall Polynomials*, Oxford University Press (Clarendon), London, New York.

M. A. NAIMARK AND D. I. STERN (1982), *Theory of Group Representations*, Springer-Verlag, New York, Berlin.

G. de B. ROBINSON (1961), *Representation Theory of the Symmetric Group*, University of Toronto Press, Toronto, Canada.

J. SAW (1977), *Zonal polynomials: An alternative approach*, J. Multivariate Anal., 7, pp. 461–467.

S. J. SCHWAGER AND B. H. MARGOLIN (1982), *Detection of multivariate normal outliers*, Ann. of Statist., 3, pp. 943–954.

A. SELBERG (1956), *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series*, J. Indian Math. Soc., 20, pp. 47–87.

R. STANLEY (1986), personal communication.

A. TAKEMURA (1984), *Zonal Polynomials*, Institute of Mathematical Statistics Lecture Notes Monograph Series, Volume 4, Hayward, CA.

A. TERRAS (1985), *Special functions for the symmetric space of positive matrices*, SIAM J. Math. Anal., 16, pp. 620–640.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

015

THE TRIDIAGONAL APPROACH TO SZEGÖ'S ORTHOGONAL POLYNOMIALS, TOEPLITZ LINEAR SYSTEMS, AND RELATED INTERPOLATION PROBLEMS*

P. DELSARTE† AND Y. GENIN†

Abstract. The basic topics of the paper are the three-term recurrence relation $x_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z)x_k(z) - zx_{k-1}(z)$ and the associated tridiagonal matrix. This relation, which underlies the Bistritz stability test, can be used as a starting point for a novel approach to the trigonometric moment problem and its relatives. In particular, the "tridiagonal approach" is shown to provide a new solution method for the classical Carathéodory–Fejér and Nevanlinna–Pick interpolation problems. The results include some Levinson-type and Schur-type algorithms, of reduced complexity, for computing reflection coefficients associated with nonnegative definite Hermitian Toeplitz matrices.

Key words. three-term recurrence, Szegö polynomials, Toeplitz matrices, interpolation problems

AMS (MOS) subject classifications. 30E05, 42A70, 30D50

1. Introduction. The Schur–Cohn test to check polynomial stability [18], [23] and the Levinson algorithm to solve Toeplitz linear systems [14], [21] are intimately related topics which have found various applications in the areas of discrete systems analysis, digital signal processing and linear least-squares estimation [19], [20], [22], [24], among others. These methods are quite efficient from the points of view of computational complexity and numerical accuracy. From a theoretical viewpoint, they owe a significant part of their popularity to the fact that they are direct implementations of two standard results of the Szegö theory of orthogonal polynomials on the unit circle [2], [12], [15], [26].

The Bistritz stability test [3], [4], [8] and the split Levinson algorithm [5], [6] have been proposed recently as substitutes for the Schur–Cohn test and the Levinson algorithm. These two methods are based on a remarkable three-term recurrence relation (with two very different interpretations). Their computational complexity and memory requirement are smaller than those of the corresponding standard algorithms; the number of multiplications and the storage space are reduced approximately by a factor 2. Although the recurrence relation just mentioned is not classical in the framework of Szegö's theory, it can actually be derived from the well-known recurrence relation for orthogonal polynomials on the unit circle [5].

This paper contains a thorough study of a suitable mathematical environment of the three-term recurrence relation or, equivalently, the associated tridiagonal matrix, underlying the Bistritz test and the split Levinson algorithm, in the general case of complex data. Our approach is of a function-theoretic nature essentially, in the sense that it brings out a new mathematical framework for some important aspects of the theory of Carathéodory functions (C-functions). In particular, the celebrated Carathéodory–Fejér (CF) and Nevanlinna–Pick (NP) interpolation problems [2], [25], which have interesting applications in systems, circuit and signal theory (see [7], [10], [11], [16], [17], for example), can be treated successfully in this framework. The "tridiagonal approach" yields not only new solvability criteria but also new efficient recursive "Schur-type" and "Nevanlinna-type" algorithms for the CF and NP problems. It is interesting to note that, in contrast with the classical Schur and Nevanlinna

* Received by the editors June 23, 1986; accepted for publication June 10, 1987.
† Philips Research Laboratory Brussels, Av. Van Becelaere 2, Box 8, B-1170 Brussels, Belgium.

algorithms that essentially involve Schur functions [2], the proposed new methods are specifically tailored to deal with Carathéodory functions. Furthermore, while establishing these results, we obtain a simple extension of the split Levinson algorithm to the complex case, as well as a straightforward derivation of a generalized version of the Bistritz stability test.

The technical contents of the paper can be summarized as follows. The tridiagonal polynomial matrix $J_n(z)$ associated with the three-term recurrence relation $x_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z)x_k(z) - zx_{k-1}(z)$, with $k = 0, 1, \cdots, n$, is introduced and examined in § 2. Two independent solution polynomials $x_k(z)$, denoted by $p_k(z)$ and $q_k(z)$, are identified in terms of certain subdeterminants of $J_n(z)$; they are characterized by their symmetry and antisymmetry properties; they are called first-kind and second-kind polynomials. In particular, all polynomials $q_k(z)$ vanish at the point $z = 1$. The theory is essentially restricted to the case where the tridiagonal matrix $J_n(1)$ is nonnegative definite.

In § 3, the sequences of such polynomials $p_k(z)$ and $q_k(z)$ are shown to be in one-to-one correspondence with well-defined sequences of first-kind and second-kind Szegö orthogonal polynomials. In that context, $p_k(z)$ and $q_k(z)$ can be identified as "singular Szegö polynomials." The explicit relation between the recurrence parameters α_k and the Schur–Cohn parameters (or reflection coefficients) of the associated Szegö polynomials is explained in detail.

The quasi-orthogonality properties of the polynomials $p_k(z)$ with respect to any positive measure underlying Szegö's theory are examined in § 4. The Gram matrix of these polynomials is shown to coincide with the tridiagonal matrix $J_n(1)$, which yields an explicit congruence relation between $J_n(1)$ and the Toeplitz matrix relative to the Szegö polynomials. A Levinson-type algorithm to compute prediction filters and reflection coefficients for nonnegative definite Hermitian Toeplitz matrices is obtained in that context.

The properties of the zeros of the polynomials $p_k(z)$ and $q_k(z)$ are discussed in § 5. It is shown that these zeros are located on the unit circle $|z| = 1$. Moreover, the zeros of $p_k(z)$ alternate with those of $q_k(z)$ and with those of $(1-z)p_{k-1}(z)$. On this occasion one gives an explanation of the Bistritz stability test for complex polynomials which has the remarkable feature of revealing the close connection between this test and the Schur–Cohn test in a transparent manner.

In § 6, the classical "coefficient problems" for C-functions are revisited in the light of the tridiagonal approach. First, a new decomposition principle for an arbitrary C-function $f(z)$ is obtained; it is based on the extraction of a lossless rational function of degree one from $f(z)$, having its pole in $z = 1$ and assuming the same value as $f(z)$ in $z = 0$. It is then shown that the CF interpolation problem with $n+1$ constraints can be solved by n iterations (at most) of this decomposition method. The new algorithm generates a sequence of complex numbers α_k from the data. With the exception of some pathological situations, the nonnegative definiteness of the associated tridiagonal matrix $J_n(1)$ turns out to be a solvability criterion for the problem. Furthermore, the general solution can be parametrized in terms of an "almost arbitrary" C-function by means of a simple homographic transformation involving the first-kind and second-kind polynomials $p_n(z)$, $p_{n+1}(z)$ and $q_n(z)$, $q_{n+1}(z)$. A suitable implementation of this solution method provides an efficient Schur-type algorithm to compute the reflection coefficients for a given nonnegative definite Toeplitz matrix; it generalizes the split Schur algorithm [6] to the complex case. It involves about twice less multiplications than the classical Schur algorithm and the same number of additions. (Thus the gain in complexity is roughly equal to that of the Levinson-type algorithm of § 5 with respect to the classical Levinson algorithm.)

The Nevanlinna-Pick interpolation problem is considered in § 7. It is shown that the methods and results of the preceding section can be carried over to this problem via suitable iterative transformations of the unit disk, mapping the given interpolation points to the origin. This gives rise to a new solvability criterion for the NP problem, together with a new explicit description of the solution space and a new Nevanlinna-type algorithm for its actual computation. As an application, a generalization of the Bistritz stability test is proposed; it involves the values assumed by the given polynomial at m points arbitrarily selected in the unit disk, where m is the polynomial degree.

2. Tridiagonal matrices and associated polynomials. Let there be given a sequence of $n+1$ complex numbers $\alpha_0, \alpha_1, \cdots, \alpha_n$. For an integer k, with $0 \leqq k \leqq n$, define the tridiagonal matrix, or Jacobi matrix,

$$J_k(z) = \begin{bmatrix} \alpha_0 + \bar{\alpha}_0 z & z & & & & \\ 1 & \alpha_1 + \bar{\alpha}_1 z & z & & & \\ & 1 & & \cdots & & \\ & & \cdots & & z & \\ & & & 1 & \alpha_k + \bar{\alpha}_k z & \end{bmatrix}, \tag{2.1}$$

where z is a complex variable and the bar denotes the conjugate. Note that (2.1) can be written in the form $J_k(z) = A_k + z\tilde{A}_k$ where A_k is the lower-triangular matrix having $\alpha_0, \alpha_1, \cdots, \alpha_k$ on the diagonal, 1 just below it and 0 elsewhere, and the tilde denotes the conjugate transpose. With $J_{k-1}(z)$ let us associate the polynomial $p_k(z)$, of formal degree k, defined by

$$p_k(z) = \det J_{k-1}(z), \tag{2.2}$$

for $k = 0, 1, \cdots, n+1$, with the convention $p_0(z) = 1$. Since $J_k(z)$ equals $z\tilde{J}_k(1/\bar{z})$ it is clear that $J_k(z)$ enjoys the *symmetry property*

$$\hat{p}_k(z) = p_k(z), \tag{2.3}$$

with the notation $\hat{x}_k(z) = z^k \bar{x}_k(1/\bar{z})$. Furthermore, computing the determinant of $J_k(z)$ by Laplace's rule, we obtain the *three-term recurrence relation*

$$p_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z)p_k(z) - zp_{k-1}(z), \tag{2.4}$$

with the initial conditions $p_{-1}(z) = 0$ and $p_0(z) = 1$. As a straightforward consequence of (2.4) we have

$$p_{k+1}(0) = \alpha_k p_k(0). \tag{2.5}$$

It is worth mentioning that (2.4) can be viewed as a special case of the Frobenius relation occurring in Padé approximation (see [13] in that respect). However, the main applications of (2.4) treated in this paper are quite different from those considered in classical Padé theory. Let us incidentally point out that the complex symmetric polynomial $p_k(z)$ yields the real trigonometric polynomial

$$\psi_k(\theta) = e^{-ik\theta} p_k(e^{2i\theta}), \tag{2.6}$$

which is a linear combination of the functions $\cos l\theta$ and $\sin l\theta$ with $l \leqq k$ and $l \equiv k \pmod 2$. The recurrence relation (2.4) assumes the nice form

$$\psi_{k+1}(\theta) = (x_k \cos\theta + y_k \sin\theta)\psi_k(\theta) - \psi_{k-1}(\theta), \tag{2.7}$$

with $x_k = 2 \operatorname{Re} \alpha_k$ and $y_k = 2 \operatorname{Im} \alpha_k$. The "real case" $y_k = 0$ (for all k) is closely related to the theory of orthogonal polynomials on the interval $[-1, +1]$. (See [26], and especially [5] in the context of this paper.)

In the sequel we exclusively consider the situation where the real symmetric matrix $J_n(1)$ *is nonnegative definite.* (Equivalently, the matrix iA_n is dissipative.) In view of the tridiagonal structure (2.1), this implies that $J_{n-1}(1)$ is positive definite. In particular, the real part of each α_k has to be positive. Our assumption can be expressed in terms of the polynomials (2.2) by the conditions $p_{n+1}(1) \geqq 0$ and $p_k(1) > 0$ for $k = 1, 2, \cdots, n$. In this context it is quite natural to introduce the *Jacobi parameters*

$$\lambda_k = p_k(1)/p_{k-1}(1), \tag{2.8}$$

for $k = 0, 1, \cdots, n+1$, with the convention $\lambda_0 = \infty$. (It is interesting to compare (2.8) with the expression $\alpha_{k-1} = p_k(0)/p_{k-1}(0)$.) The *positivity constraints* above can then be written in the form

$$\lambda_k > 0 \quad \text{for } 1 \leqq k \leqq n \quad \text{and } \lambda_{n+1} \geqq 0. \tag{2.9}$$

The Jacobi parameters λ_k can be determined from the data by means of a simple continued fraction. Indeed, (2.4) yields the recurrence relation

$$\lambda_{k+1} = 2 \operatorname{Re} \alpha_k - \lambda_k^{-1} \quad \text{for } 0 \leqq k \leqq n. \tag{2.10}$$

Without going into any detail let us mention that the general situation where $J_n(1)$ has an arbitrary inertia could presumably be treated with the help of the theory of pseudo-Carathéodory functions developed in [9].

For certain applications it is useful to consider a dual family of solutions of the recurrence relation (2.4). The *second-kind polynomials* $q_k(z)$ are defined by

$$q_k(z) = (1 - z) \det J_{k-1}^0(z), \tag{2.11}$$

where $J_k^0(z)$ is the submatrix of $J_k(z)$ obtained by deleting its first row and column. It is clear that we actually have

$$q_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z) q_k(z) - z q_{k-1}(z), \tag{2.12}$$

for $k = 1, 2, \cdots, n$, with the initial conditions $q_0(z) = 0$ and $q_1(z) = 1 - z$. The main difference with the first-kind polynomials $p_k(z)$ lies in the fact that $q_k(z)$ enjoys the *antisymmetry property*

$$\hat{q}_k(z) = -q_k(z), \tag{2.13}$$

instead of (2.3). Furthermore, all polynomials $q_k(z)$ vanish at the point $z = 1$. As a consequence of the assumption (2.9) we can show that the derivative of $q_k(z)$ is negative at $z = 1$, for $1 \leqq k \leqq n+1$. Applying the classical Jacobi theorem [1] to the corner entries of $J_k(z)$ we obtain the interesting relation

$$p_k(z) q_{k+1}(z) - q_k(z) p_{k+1}(z) = (1 - z) z^k, \tag{2.14}$$

by use of (2.2) and (2.11). Equivalently, (2.14) can be deduced from (2.4) and (2.12), by induction. Note that we have $q_k(0) = \alpha_0^{-1} p_k(0)$ for $k \geqq 1$.

Let us add a short comment on the role played by the point $z = 1$ in the whole theory (see (2.8) and (2.11), for example). In our approach it is quite important that the second-kind polynomials $q_k(z)$ all vanish for *some* fixed point ζ of unit modulus and that the first-kind polynomials satisfy the inequality $\zeta^{-k/2} p_k(\zeta) \geqq 0$, with possible equality when $k = n+1$ only. The polynomials in question have the general form $p_k(z) = \det J_{k-1}(z)$ and $q_k(z) = (\zeta^{1/2} - \zeta^{-1/2} z) \det J_{k-1}^0(z)$, with $J_k(z)$ as in (2.1). In this paper we make the choice $\zeta = 1$ for the sake of definiteness and simplicity, which entails no loss of generality. Indeed, for an arbitrary ζ with $|\zeta| = 1$, we can construct "normalized polynomials" $p_k'(z) = \zeta^{-k/2} p_k(\zeta z)$ and $q_k'(z) = \zeta^{-k/2} q_k(\zeta z)$ having the

required properties (with $\zeta' = 1$); the corresponding parameters of the tridiagonal matrix (2.1) are deduced from the original parameters by the formula $\alpha'_k = \alpha_k \zeta^{-k/2}$.

3. Connections with the Szegö polynomials. From the symmetric polynomials $p_{k+1}(z)$ and $p_k(z)$ let us construct the *comonic polynomial* $a_k(z)$ of formal degree k via the identity

$$p_{k+1}(0)(1-z)a_k(z) = p_{k+1}(z) - \lambda_{k+1} z p_k(z), \tag{3.1}$$

where λ_k is the Jacobi parameter (2.8). Let $\rho_k = a_{k,k}$ denote the coefficient of z^k in $a_k(z)$. By definition, $\bar{p}_{k+1}(0) + \rho_k p_{k+1}(0) = \lambda_{k+1}\bar{p}_k(0)$. Using (2.5) and (2.10) we deduce

$$\rho_k = \left(1 - \frac{1}{\lambda_k \alpha_k}\right) \frac{\bar{p}_k(0)}{p_k(0)}, \tag{3.2}$$

which yields the remarkable identity

$$|\alpha_k|^2 (1 - |\rho_k|^2) = \lambda_k^{-1} \lambda_{k+1}. \tag{3.3}$$

Therefore, the positivity conditions (2.9) can be expressed by $|\rho_k| < 1$ for $1 \leqq k \leqq n-1$ and $|\rho_n| \leqq 1$.

Let $\hat{a}_k(z) = z^k \bar{a}_k(1/\bar{z})$ denote the reciprocal of $a_k(z)$. As explained below, the monic polynomials $\hat{a}_k(z)$ constitute a family of *Szegö polynomials*. To discover the intrinsic meaning of the parameter ρ_k let us compute the polynomial $a_{k-1}(z) + \rho_k \hat{z} a_{k-1}(z)$ by using (3.1) with k replaced by $k-1$. Applying (2.4), (2.10), and (3.2) we readily obtain the recurrence relation

$$a_k(z) = a_{k-1}(z) + \rho_k z \hat{a}_{k-1}(z). \tag{3.4}$$

Therefore, the numbers ρ_k constitute the sequence of *Schur-Cohn parameters* [18], [23] of the polynomial $a_n(z)$.

Some "second-kind Szegö polynomials" can be obtained from the antisymmetric polynomials (2.11) by a relation similar to (3.1). Define the polynomial $r_k(z)$, of formal degree k, via the identity

$$p_{k+1}(0)(1-z)r_k(z) = q_{k+1}(z) - \lambda_{k+1} z q_k(z), \tag{3.5}$$

for $k = 0, 1, \cdots, n$. Note that we have $r_k(0) = \alpha_0^{-1}$ and $r_{k,k} = -\bar{\alpha}_0^{-1}\rho_k$ for all k. By a similar argument as above we can prove the recurrence relation

$$r_k(z) = r_{k-1}(z) - \rho_k z \hat{r}_{k-1}(z). \tag{3.6}$$

This shows that the polynomials $\hat{r}_k(z) = z^k \bar{r}_k(1/\bar{z})$ defined from (3.5) constitute the family of second-kind Szegö polynomials associated with the first-kind Szegö polynomials $\hat{a}_k(z)$ (see [12]).

Next, let us explain how the polynomials $p_k(z)$ and $q_k(z)$ can be recovered from the Szegö polynomials $\hat{a}_k(z)$ and $\hat{r}_k(z)$. Using (3.1) and (3.5) we readily derive the relations

$$\begin{aligned} \lambda_{k+1} p_k(z) &= p_{k+1}(0) a_k(z) + \bar{p}_{k+1}(0) \hat{a}_k(z), \\ \lambda_{k+1} q_k(z) &= p_{k+1}(0) r_k(z) - \bar{p}_{k+1}(0) \hat{r}_k(z), \end{aligned} \tag{3.7}$$

for $k = 0, 1, \cdots, n$. Alternative relations, involving a_{k-1} and r_{k-1} instead of a_k and r_k, can be obtained by substituting (3.4) and (3.6) into (3.7). The results are

$$\begin{aligned} p_k(z) &= p_k(0) a_{k-1}(z) + \bar{p}_k(0) z \hat{a}_{k-1}(z), \\ q_k(z) &= p_k(0) r_{k-1}(z) - \bar{p}_k(0) z \hat{r}_{k-1}(z), \end{aligned} \tag{3.8}$$

for $k=1, 2, \cdots, n+1$. Thus, within a constant factor, $p_k(z)$ and $q_k(z)$ coincide with the "singular polynomials" (3.4) and (3.6) obtained by substituting the number

$$\varepsilon_k = \bar{p}_k(0)/p_k(0), \tag{3.9}$$

of unit modulus, for the Schur-Cohn parameter ρ_k. Since $q_k(1)$ vanishes, the second relation (3.8) yields the useful result

$$\varepsilon_k = r_{k-1}(1)/\bar{r}_{k-1}(1). \tag{3.10}$$

This allows us to determine the sequence of numbers ε_k from the parameters ρ_k. Indeed, (3.6) yields the recurrence relation

$$\varepsilon_{k+1} = (\varepsilon_k - \rho_k)/(1 - \varepsilon_k\bar{\rho}_k). \tag{3.11}$$

The initial value is $\varepsilon_1 = \bar{\alpha}_0/\alpha_0$. Equivalently, (3.11) can be written in the form of the factorization

$$1 - |\rho_k|^2 = (1 - \varepsilon_k\bar{\rho}_k)(1 + \varepsilon_{k+1}\bar{\rho}_k). \tag{3.12}$$

(Note that the theory is significantly simpler in the case of real data, which yields $\varepsilon_k = 1$ for all k.)

It remains to find out an expression for the values $p_k(0)$ or, equivalently, the numbers α_k, in terms of the Schur-Cohn parameters ρ_k. From (3.2), (3.3), and (3.12) we deduce both identities $\alpha_k(1 - \bar{\varepsilon}_k\rho_k) = \lambda_k^{-1}$ and $\alpha_k(1 + \varepsilon_{k+1}\bar{\rho}_k) = \lambda_{k+1}$, whence

$$\alpha_{k-1}\alpha_k = (1 + \varepsilon_k\bar{\rho}_{k-1})^{-1}(1 - \bar{\varepsilon}_k\rho_k)^{-1}. \tag{3.13}$$

For a given value of α_0, this allows us to determine the α_k's from the ρ_k's, with the convention $\rho_0 = 1$. In summary, the sequence of symmetric polynomials $p_k(z)$ can be computed from the Szegö polynomials $\hat{a}_k(z)$ by means of the first formula (3.8), with the help of (3.13) and (2.5). Furthermore, the recurrence relation (2.4), with the appropriate numbers α_k, can be deduced from the Szegö relation (3.4). A verification of the latter property is left to the reader. There are completely similar results concerning the associated antisymmetric polynomials $q_k(z)$. This subject will be examined in further detail in the next section, from the viewpoint of Toeplitz systems of linear equations.

In the case where $J_n(1)$ is singular we have $\lambda_{n+1} = 0$ (i.e., $\rho_n = \varepsilon_{n+1}$), which yields the identities $p_{n+1}(z) = p_{n+1}(0)(1-z)a_n(z)$ and $q_{n+1}(z) = p_{n+1}(0)(1-z)r_n(z)$ via (3.1) and (3.5). Since $q_{n+1}(z)$ has a simple zero at $z=1$, the value of $r_n(1)$ cannot vanish; this is the only restriction that our approach (with the choice $\zeta = 1$) places on the Szegö theory.

Let us finally comment on the environment of the formulas (3.8). The two-variable *Christoffel-Darboux polynomial* of the second kind can be defined by

$$Q_k(\zeta, z) = \zeta\hat{r}_{k-1}(\zeta)r_{k-1}(z) - r_{k-1}(\zeta)z\hat{r}_{k-1}(z) \tag{3.14}$$

(see [26]). In view of (3.10) we have $q_k(z) = \beta_k Q_k(1, z)$ for a real constant β_k. There is a similar interpretation for $p_k(z)$ in terms of a suitable two-variable "Green polynomial" $P_k(\zeta, z)$.

4. Orthogonality relations and Levinson-type algorithm. A family of monic polynomials $\hat{a}_k(z)$ satisfying the recurrence relation (3.4) with the constraints $|\rho_k| < 1$ for $1 \leqq k \leqq n-1$ and $|\rho_n| \leqq 1$ is known to be a Szegö family, in the sense that the $\hat{a}_k(z)$ are pairwise orthogonal on the unit circle with respect to a certain positive measure $d\omega(\theta)$.

Let us explain this property in precise terms (see [2], [12], [26]). The inner product $\langle x, y\rangle$ of any two pseudopolynomials $x(z)$ and $y(z)$ with respect to $d\omega(\theta)$ is defined by

$$\langle x, y\rangle = \int_0^{2\pi} \bar{x}(e^{i\theta}) y(e^{i\theta})\, d\omega(\theta). \tag{4.1}$$

For an appropriate choice of the measure, the polynomials $\hat{a}_k(z)$ satisfy the *orthogonality relations*

$$\langle \hat{a}_k, \hat{a}_l\rangle = \sigma_k \delta_{k,l}, \tag{4.2}$$

for $0 \leqq k, l \leqq n$, where δ is the Kronecker symbol and σ_k is a nonnegative real number. More precisely, we have $\sigma_k > 0$ for $0 \leqq k \leqq n-1$ and $\sigma_n \geqq 0$, with $\sigma_n = 0$ if and only if $|\rho_n| = 1$. In fact, the *squared norms* σ_k obey the recurrence relation $\sigma_k = (1-|\rho_k|^2)\sigma_{k-1}$. For the sake of normalization it proves convenient to set

$$\sigma_0 = c_0 = \mathrm{Re}\,(\alpha_0^{-1}). \tag{4.3}$$

From (3.3) and (2.5) we deduce the useful identity

$$\lambda_k = 2\sigma_{k-1}|p_k(0)|^2. \tag{4.4}$$

Except in the *singular case* $|\rho_n| = 1$ (i.e., $\lambda_{n+1} = 0$), the measure $d\omega(\theta)$ is not unique. However its first $2n+1$ trigonometric moments

$$c_s = \int_0^{2\pi} e^{-is\theta}\, d\omega(\theta), \qquad -n \leqq s \leqq n, \tag{4.5}$$

are uniquely determined from $a_n(z)$ or, equivalently, from the parameters $\rho_1, \cdots, \rho_n$. (Here c_0 is supposed to be given. Note the property $c_{-s} = \bar{c}_s$.) Conversely, the polynomials $a_k(z)$ can be computed from the moments (4.5) as follows. Construct the *Hermitian Toeplitz matrix*

$$C_k = [c_{s-t} : 0 \leqq s, t \leqq k], \tag{4.6}$$

which is the Gram matrix of the monomials $1, z, \cdots, z^k$ with respect to the inner product (4.1). It is positive definite for $0 \leqq k \leqq n-1$ and nonnegative definite for $k = n$. The orthogonality relations (4.2) can be expressed by the fact that the coefficient vector $a_k = [a_{k,t} : 0 \leqq t \leqq k]^T$ of the comonic polynomial $a_k(z) = \sum_{t=0}^{k} a_{k,t} z^t$ is the solution of the system of linear equations

$$C_k a_k = [\sigma_k, 0, \cdots, 0]^T. \tag{4.7}$$

Note that we have $\sigma_k = \det C_k / \det C_{k-1}$ for $k = 1, 2, \cdots, n$.

In view of (3.7), the coefficient vector $p_k = [p_{k,t} : 0 \leqq t \leqq k]^T$ of the symmetric polynomial $p_k(z) = \sum_{t=0}^{k} p_{k,t} z^t$ satisfies the linear system

$$C_k p_k = [\bar{\tau}_k, 0, \cdots, 0, \tau_k]^T, \tag{4.8}$$

for $k \geqq 1$, where $\tau_k = \sigma_k \bar{p}_{k+1}(0)/\lambda_{k+1} = (2p_{k+1}(0))^{-1}$, by (4.4). Applying (2.5) we obtain the important formula

$$\alpha_k = \tau_{k-1}/\tau_k, \tag{4.9}$$

with the convention $\tau_0 = (2\alpha_0)^{-1}$. From (4.8) we then deduce the set of relations

$$\begin{aligned} \langle p_k, p_k\rangle &= \mathrm{Re}\,(\alpha_k^{-1}), \\ \langle p_k, p_l\rangle &= (2\alpha_l \alpha_{l+1} \cdots \alpha_k)^{-1} \quad \text{for } k > l, \\ \langle p_k, z^t p_l\rangle &= 0 \quad \text{for } k - l > t > 0, \end{aligned} \tag{4.10}$$

for $0 \leqq k \leqq n$. Let us then introduce the pseudopolynomials $v_k(z)$ by normalizing and

shifting the polynomials $p_k(z)$ as follows:

$$\begin{aligned} v_{2t}(z) &= \sqrt{2}\alpha_{2t} z^{-t} p_{2t}(z), \\ v_{2t+1}(z) &= \sqrt{2}\,\bar{\alpha}_{2t+1}\, z^{-t} p_{2t+1}(z). \end{aligned} \tag{4.11}$$

It is easily seen that the relations (4.10) assume the simple form

$$\begin{aligned} \langle v_k, v_k \rangle &= 2\,\mathrm{Re}\,(\alpha_k), \\ \langle v_k, v_{k-t} \rangle &= \delta_{t,1} \quad \text{for } t \geqq 1. \end{aligned} \tag{4.12}$$

This shows that the tridiagonal matrix $J_n(1) = A_n + \tilde{A}_n$ can be interpreted as the Gram matrix of the pseudopolynomials $v_0(z), v_1(z), \cdots, v_n(z)$ with respect to the inner product (4.1). There is a close connection between (4.11) and (2.6). It appears that $\psi_k(\theta/2)$ and $\psi_l(\theta/2)$ are orthogonal when k and l have the same parity. This is generally not true for opposite parities, except in the case of real data (see [5]).

It is interesting to note that (4.12) can be interpreted as a congruence relation between the Toeplitz matrix C_n and the Jacobi matrix $J_n(1)$. Indeed, we have

$$J_n(1) = \tilde{V}_n C_n V_n, \tag{4.13}$$

where V_n is the square matrix of order $n+1$ whose columns are the coefficient vectors of the pseudo-polynomials $v_k(z)$. By definition, V_n is a triangular matrix within permutation of its rows. The result (4.13), or simply (4.4), yields an interesting *positivity test* for a given Hermitian Toeplitz matrix C_n of order $n+1$. Indeed, it shows that the conditions (2.9) on the Jacobi parameters λ_k characterize precisely the fact that C_n is nonnegative definite and C_{n-1} positive definite. The λ_k's can be obtained from the entries of C_n as by-products of the Levinson-type algorithm described below.

Given a nonnegative definite Toeplitz matrix C_n of nullity 0 or 1, the polynomial $a_n(z)$ defined from (4.7) yields the optimal *prediction filter* of length n for a stationary stochastic process having C_n as its autocorrelation matrix [20], [22]. The *Levinson algorithm* [14], [21] is a recursive method to compute the predictors $a_k(z)$ together with the Schur-Cohn parameters ρ_k (called reflection coefficients in this context). Let us now explain how the "singular predictors" $p_k(z)$ can be determined from C_n by a recursive method essentially different from but formally analogous to the Levinson algorithm. The outcome will be an efficient procedure to compute the desired predictor $a_n(z)$, together with the reflection coefficients ρ_k and the Jacobi parameters λ_k. The special case of a real Toeplitz matrix C_n has been recently treated in detail by the authors, who proposed the name "split Levinson algorithm" for their new method [5].

Recall that the coefficient vector of $p_k(z)$ is the solution of the Toeplitz system (4.8). Assuming $p_{k-1}(z)$, $p_k(z)$ and τ_{k-1} to be available at the kth step of the algorithm, let us indicate how to compute $p_{k+1}(z)$ and τ_k. In view of (4.8) we simply have

$$\tau_k = \sum_{i=0}^{k} c_{k-i}\, p_{k,i}. \tag{4.14}$$

Then the parameter α_k is given by (4.9) and the updated polynomial $p_{k+1}(z)$ results from the recurrence relation (2.4). The initial conditions are $p_{-1}(z) = 0$, $p_0(z) = 1$ and $\tau_0 = (2\alpha_0)^{-1}$. (Thus, $\mathrm{Re}\,\tau_0 = c_0/2$.) The numbers of real additions and multiplications in the kth step of the algorithm are found to be $8k+15$ and $4k+14$, respectively. The corresponding numbers for the Levinson algorithm are $8k-6$ and $8k-4$. The gain in computational complexity stems from the symmetry property $p_{k,k-t} = \bar{p}_{k,t}$ of the coefficients of $p_k(z)$, which allows one to significantly economize on arithmetic operations. Furthermore, the same property leads to a reduction of the memory space by a factor 2 (roughly speaking).

The predictor $a_n(z)$ can be computed at the end of the algorithm by means of (2.8) and (3.1) with $k=n$. The successive Jacobi parameters λ_k can be determined from the numbers α_k with the help of the recurrence relation (2.10). Then the reflection coefficients ρ_k are obtainable from (3.2). Note that $\varepsilon_k = \tau_{k-1}/\bar{\tau}_{k-1}$.

To conclude this section let us indicate how the entries $c_0, c_1, \cdots, c_n$ of the Hermitian Toeplitz matrix C_n can be computed from the numbers $\alpha_0, \alpha_1, \cdots, \alpha_n$ in a direct manner. As shown in § 6, we have the remarkable interpolation property

$$\frac{q_{n+1}(z)}{p_{n+1}(z)} = \alpha_0^{-1} + 2\sum_{t=1}^{n} c_t z^t + O(z^{n+1}), \tag{4.15}$$

where $p_{n+1}(z)$ and $q_{n+1}(z)$ are obtained from the recurrence relations (2.4) and (2.12) with the initializations indicated above. The significance of (4.15) will become clear in the sequel.

5. Zero location and Bistritz stability test. It is well known that the "predictor polynomial" $a_k(z)$ is devoid of zeros in the closed unit disk $|z| \leqq 1$ if and only if the associated Hermitian Toeplitz matrix C_k is positive definite (in case $c_0 > 0$). In view of (3.4), this can be interpreted as the *Schur-Cohn stability test* $|\rho_t| < 1$ for $< 1 \leqq t \leqq k$ (see [18], [23]). Furthermore, if C_n is nonnegative definite and has rank n then $a_n(z)$ has n distinct zeros on the unit circle $|z| = 1$ (see [15]). The same statements can be made concerning the polynomials $r_k(z)$.

Next, let us examine the properties of the zeros of the polynomials $p_k(z)$ and $q_k(z)$. Consider first the "nonsingular case" where $J_n(1)$ is strictly positive definite. It follows from (3.7) or (3.8) that $p_k(z)$ has k distinct zeros on the unit circle (for $1 \leqq k \leqq n+1$) and $q_k(z)$ has the same property. In fact, the zeros of $p_k(z)$ *separate* those of $q_k(z)$ on the unit circle. This follows from the fact that the quotient function $g_k(z) = q_k(z)/p_k(z)$ has degree k and is a Carathéodory function of *lossless* type, in the sense that it satisfies the inequality $\operatorname{Re} g_k(z) \geqq 0$ in the unit disk $|z| < 1$ and the equality $\operatorname{Re} g_k(z) = 0$ almost everywhere on the unit circle $|z| = 1$. Furthermore, it can be shown that the zeros of $p_k(z)$ separate those of $(1-z)p_{k-1}(z)$, because the quotient of these polynomials is a lossless function of degree k. The proof of the former result is elementary; the proof of the latter is contained in the derivation of the Bistritz test given below.

In the "singular case" where $\det J_n(1)$ vanishes and $J_{n-1}(1)$ is positive definite, the properties mentioned above have to be slightly modified (when $k = n+1$). Recall that $p_{n+1}(z)$ and $q_{n+1}(z)$ are proportional to $(1-z)a_n(z)$ and $(1-z)r_n(z)$, respectively. It turns out that the zeros of $a_n(z)$ separate those of $r_n(z)$, on the one hand, and those of $p_n(z)$, on the other hand.

The remaining part of this section is devoted to explaining how the *Bistritz stability criterion* fits nicely into the general framework of the paper. The Bistritz test is an interesting new method to check whether a given complex polynomial $x_n(z)$ of degree n is *stable*, in the sense that it does not vanish in the closed unit disk $|z| \leqq 1$ (see [3], [4], [8]); this method has a lower computational complexity than the classical Schur-Cohn stability test. Without loss of generality we assume that $x_n(1)$ is real. The main part of the Bistritz algorithm consists in computing the descending sequence of symmetric polynomials $p_k(z)$, for $k = n, n-1, \cdots, 1, 0$, from the recurrence relation (2.4) with the initialization

$$p_n(z) = x_n(z) + \hat{x}_n(z), \qquad (1-z)p_{n-1}(z) = x_n(z) - \hat{x}_n(z). \tag{5.1}$$

(Note that $p_0(z)$ is generally not equal to unity as in § 2. In the present context, the normalization is not given a priori.) The Bistritz criterion says that $x_n(z)$ is stable if

and only if the parameters $\lambda_k = p_k(1)/p_{k-1}(1)$ exist and are positive for $k = n, n-1, \cdots, 1$, which means that the tridiagonal matrix $J_{n-1}(1)$ built from the complex numbers $\alpha_k = p_{k+1}(0)/p_k(0)$ is positive definite. The following argument contains a simple proof of this result.

Using (3.7), (3.8), and (5.1) we derive the identity

$$\frac{x_n(z)+\hat{x}_n(z)}{x_n(z)-\hat{x}_n(z)} = \frac{\lambda_n}{2}\left[\frac{1+z}{1-z} + \frac{a_{n-1}(z)-\varepsilon_n \hat{a}_{n-1}(z)}{a_{n-1}(z)+\varepsilon_n \hat{a}_{n-1}(z)}\right], \tag{5.2}$$

by straightforward computation. If $J_{n-1}(1)$ is positive definite, then so is the Toeplitz matrix C_{n-1}, which implies that $a_{n-1}(z)$ is stable. Therefore, both terms in the right-hand side of (5.2) are lossless functions, which implies that $x_n(z)$ is stable. Conversely, if $x_n(z)$ is stable, then the left-hand side of (5.2) is a lossless function, having a pole at $z = 1$ with positive mass λ_n. The second term in (5.2) is precisely obtained by extraction of this pole; hence it is a lossless function, which implies that $a_{n-1}(z)$ is stable. It then follows from (3.3), via the Schur-Cohn criterion, that all parameters λ_k are positive.

6. Carathéodory–Fejér interpolation problem. This section contains a new approach to the trigonometric moment problem or, equivalently, the coefficient problem for Carathéodory functions. In fact, we are mainly interested in the partial trigonometric moment problem, which is equivalent to the Carathéodory-Fejér interpolation problem [2], [25]. Recall that a function $f(z)$ of the complex variable z belongs to the class C of *Carathéodory functions* if it is analytic and satisfies $\operatorname{Re} f(z) \geqq 0$ in the unit disk $|z| < 1$. Consider the Maclaurin expansion $f(z) = \gamma_0 + 2\sum_{t=1}^{\infty} c_t z^t$. The well-known *Carathéodory-Toeplitz theorem* says that $f(z)$ is a C-function if and only if the Hermitian Toeplitz matrix C_k built from the coefficients c_t, with $c_0 = \operatorname{Re} \gamma_0$ and $c_{-t} = \bar{c}_t$, is nonnegative definite for all k.

The Schur criterion provides a very interesting alternative solution to the same problem. Recall that a function $\phi(z)$ belongs to the class S of *Schur functions* if it is analytic and satisfies $|\phi(z)| \leqq 1$ in the unit disk. From a given function $\phi(z)$ let us construct a sequence of functions $\phi_k(z)$ by means of the Schur recurrence relation

$$\phi_{k+1}(z) = \frac{\phi_k(z) - \phi_k(0)}{z(1 - \bar{\phi}_k(0)\phi_k(z))}, \tag{6.1}$$

for $k = 0, 1$, etc., with the initialization $\phi_0(z) = \phi(z)$. The *Schur criterion* says that $\phi(z)$ is an S-function if and only if we have $|\phi_k(0)| \leqq 1$ for all k. (In case $|\phi_n(0)| = 1$ for a certain n, the criterion says that $\phi_n(z)$ has to be a constant; this characterizes the Blaschke products $\phi(z)$ of degree n.) Since the bilinear transform $\phi(z) = (1 - f(z))/(1 + f(z))$ establishes a bijection between the class S and the class C (supplemented with the "function" $f(z) = \infty$), the Schur criterion can be used to solve the Carathéodory coefficient problem. Note the identity $\phi_k(0) = \rho_k$, for $k \geqq 1$, where the ρ_k's are the reflection coefficients defined as in §§ 3 and 4.

Let there be given $n+1$ complex numbers $\gamma_0, c_1, \cdots, c_n$. The *Carathéodory-Fejér* (CF) *interpolation problem* requires to determine the set of C-functions $f(z)$ satisfying

$$f(z) = \gamma_0 + 2\sum_{t=1}^{n} c_t z^t + O(z^{n+1}). \tag{6.2}$$

This problem can be completely solved with the help of the Schur relation (6.1). When the Toeplitz matrix $C_n = [c_{s-t} : 0 \leqq s, t \leqq n]$, with $c_0 = \operatorname{Re} \gamma_0$ and $c_{-t} = \bar{c}_t$, is positive definite, there is an infinite number of solutions $f(z)$, given by

$$f(z) = \frac{r_n(z) - z\phi(z)\hat{r}_n(z)}{a_n(z) + z\phi(z)\hat{a}_n(z)}, \tag{6.3}$$

where $\phi(z)$ is an arbitrary S-function. (Here $a_n(z)$ and $r_n(z)$ denote the reciprocals of the first-kind and second-kind Szegö polynomials associated with C_n. The parameter α_0 is set equal to γ_0^{-1}.) When C_n is nonnegative definite but singular, there is a unique candidate solution, namely the rational lossless function $f(z) = r_m(z)/a_m(z)$ where m is the rank of C_n. In the remaining case, there is no solution. These results are quite classical (see especially [2]), except perhaps the explicit representation (6.3), for which a detailed proof can be found in [10].

In the sequel it is explained how the symmetric and antisymmetric polynomials $p_k(z)$ and $q_k(z)$ or, equivalently, their recurrence parameters α_k, yield alternative solution methods for the problems mentioned above. Roughly speaking, these data replace the classical polynomials $a_k(z)$ and $r_k(z)$, and the reflection coefficients ρ_k.

Given a C-function $g(z)$ let us denote by $\mu(g)$ the *inverse of the mass of* $g(z)$ *at the point* $z = 1$. More precisely,

$$\mu(g) = [\lim_{z \uparrow 1} (1-z)g(z)]^{-1}. \tag{6.4}$$

Thus we have $\mu(g) = \infty$ when $g(z)$ has no quasi-pole at $z = 1$. By convention, we set $\mu(g) = 0$ for the trivial "function" $g(z) = \infty$. In the remaining cases, $\mu(g)$ is a positive real number. The following lemma plays a crucial role in our approach; it could be viewed as a substitute for the Schwarz lemma underlying the Schur criterion.

LEMMA 1. *Let* $f_k(z)$ *be a* C*-function and define* $f_{k+1}(z)$ *by means of the relation*

$$f_k(z) = \frac{\alpha_k + \bar{\alpha}_k z}{1-z} + \frac{1}{(1-z)(1-z^{-1})f_{k+1}(z)}, \tag{6.5}$$

where $\alpha_k = f_k(0)$. *Then* $f_{k+1}(z)$ *is a* C*-function satisfying*

$$0 \leqq \mu(f_{k+1}) \leqq 2 \operatorname{Re} \alpha_k, \tag{6.6}$$

with the possibility $f_{k+1}(z) = \infty$. *Conversely, let* α_k *be a complex number and let* $f_{k+1}(z)$ *be a* C*-function, subject to the constraint* (6.6). *Then the right-hand side of* (6.5) *is a* C*-function satisfying* $f_k(0) = \alpha_k$.

Proof. Set $\psi_k(z) = (f_k(z) - \alpha_k)/z(f_k(z) + \bar{\alpha}_k)$. We readily verify that (6.5) can be written in the form

$$f_{k+1}(z) = \frac{1}{4 \operatorname{Re} \alpha_k} \left[\frac{1+z}{1-z} + \frac{1+\psi_k(z)}{1-\psi_k(z)} \right]. \tag{6.7}$$

Assume first $f_k(z)$ to be a C-function, with $f_k(0) = \alpha_k$. It follows from the Schwarz lemma that $\psi_k(z)$ is an S-function. Hence, (6.7) shows that $f_{k+1}(z)$ is a C-function (since it is the sum of two C-functions). Furthermore, (6.5) yields the identity

$$\mu^{-1}(f_k) + \mu(f_{k+1}) = 2 \operatorname{Re} \alpha_k, \tag{6.8}$$

which proves the bound (6.6).

Next, consider a C-function $f_{k+1}(z)$ satisfying (6.6), for a given number α_k. The interpolation property $f_k(0) = \alpha_k$ is a straightforward consequence of (6.5). Furthermore, the constraint (6.6) exactly says that the mass of the first summand of (6.7) at $z = 1$ does not exceed that of $f_{k+1}(z)$. Therefore, the second summand is a C-function, which means that $\psi_k(z)$ is an S-function, implying that $f_k(z)$ is a C-function. □

It is worth mentioning that Lemma 1 can be deduced immediately from the theory of pseudo-Carathéodory functions [9]. The first summand of (6.5), denoted here by $u_k(z)$, is the lossless function of degree 1, with its pole in $z = 1$, satisfying $u_k(0) = f_k(0)$. Then the second summand has the form of the canonical factorization of the residual

pseudo-Carathéodory function $f_k(z)-u_k(z)$; indeed, the inverses of the functions $(1-z)(1-z^{-1})$ and $f_{k+1}(z)$ are the density factor and the Carathéodory factor of $f_k(z)-u_k(z)$ (see [9]).

The recurrence relation (6.5) can be used to solve the CF interpolation problem (6.2) by an iterative procedure formally similar to the Schur method based on (6.1). Indeed, the expression (6.5) with $k=0$ provides a parametrization of all C-functions satisfying the interpolation constraint $f_0(0)=\alpha_0$ in terms of a C-function $f_1(z)$ subject to the only restriction $0\leqq\mu(f_1)\leqq 2\ \mathrm{Re}\ \alpha_0$. The remaining interpolation constraints can be transferred to the function $f_1(z)$ and the method above can be iterated in an obvious way. Thus it is seen that the problem admits a solution if and only if the inequalities (6.6) are satisfied at each step of the algorithm. Furthermore, it is intuitively clear that there is an infinite number of solutions unless one of the bounds (6.6) is tight.

Let us now examine the method in some detail. The recurrence relation (6.5) is used to solve the problem (6.2) with the initialization $f_0(z)=f^{-1}(z)$. From a computational viewpoint it is interesting to express (6.5) in terms of the functions $w_k(z)$ defined from the identity

$$f_k(z)=w_{k-1}(z)/(1-z)w_k(z), \tag{6.9}$$

with $w_{-1}(z)=1-z$ and $w_0(z)=f(z)$. Thus we have

$$w_k(z)=[(1-z)^k f_0(z)f_1(z)\cdots f_k(z)]^{-1}. \tag{6.10}$$

Using (6.9) we can write (6.5) in the form

$$zw_{k+1}(z)=(\alpha_k+\bar{\alpha}_k z)w_k(z)-w_{k-1}(z). \tag{6.11}$$

It is easily seen that, except in some "pathological situations," the interpolation constraints (6.2) determine the $n+1$ parameters $\alpha_0,\alpha_1,\cdots,\alpha_n$ (and conversely). Indeed, from (6.9) and (6.11) we deduce the identities

$$\alpha_k=w_{k-1,0}/w_{k,0}, \tag{6.12}$$

$$w_{k+1,t}=\alpha_k w_{k,t+1}+\bar{\alpha}_k w_{k,t}-w_{k-1,t+1}, \tag{6.13}$$

with $w_k(z)=\sum_{t=0}^{\infty}w_{k,t}z^t$. In general, these have to be used for $0\leqq k\leqq n$ and $0\leqq t\leqq n-k-1$, with the initial conditions $w_{-1,0}=1$, $w_{-1,1}=-1$, $w_{-1,t}=0$ for $t\geqq 2$, and $w_{0,0}=\gamma_0$, $w_{0,t}=2c_t$ for $t\geqq 1$. But the algorithm is bound to stop when it meets the situation $w_{l,0}=0$ for a certain l with $0\leqq l\leqq n$. If $w_{l,t}=0$ for $t=0,1,\cdots,n-l-1$, then the interpolation constraints are satisfied by the choice $w_l(z)=0$, i.e., $f_l(z)=\infty$; this will be referred to as the *degenerate case*. If $w_{l,t}\neq 0$ for some t, then it is clear that the CF problem admits no solution. The case where $w_{k,0}\neq 0$ for $0\leqq k\leqq n$ will be called *nondegenerate*; it will be appropriately characterized by the convention $l=n+1$ (which does not refer to the properties of $w_{n+1}(z)$).

From the parameters $\alpha_0,\alpha_1,\cdots,\alpha_{l-1}$ thus obtained let us construct the tridiagonal matrices $J_k(z)$, as in (2.1), and the corresponding polynomials $p_k(z)$ and $q_k(z)$. We are now in a position to describe the complete solution of the CF problem in terms of these data. (Henceforth we rule out the case where $w_{l,0}=0$ and $w_{l,t}\neq 0$ for some $t\geqq 1$.)

THEOREM 2. *A necessary and sufficient condition for the* CF *interpolation problem to be solvable is that the matrix $J_{l-1}(1)$ be nonnegative definite. In the degenerate case ($l\leqq n$) there is a unique solution, given by $f(z)=q_l(z)/p_l(z)$; it is lossless and its degree equals l or $l-1$ according as $J_{l-1}(1)$ is positive definite or singular. In the nondegenerate case ($l=n+1$), if $J_n(1)$ is singular then there is a unique solution, namely the rational*

lossless function $f(z)=q_{n+1}(z)/p_{n+1}(z)$, *which has degree* n; *if* $J_n(1)$ *is positive definite then there is an infinite number of solutions* $f(z)$, *given by*

$$f(z)=\frac{zq_n(z)-(1-z)g(z)q_{n+1}(z)}{zp_n(z)-(1-z)g(z)p_{n+1}(z)}, \tag{6.14}$$

where $g(z)$ *is a C-function subject to the only constraint*

$$0\leqq\mu(g)\leqq\lambda_{n+1}. \tag{6.15}$$

Proof. Let us write the recurrence relation (6.11) in the form

$$J_k(z)[w_0(z),-w_1(z),\cdots,(-1)^k w_k(z)]^T=[w_{-1}(z),0,\cdots,0,(-1)^k z w_{k+1}(z)]^T, \tag{6.16}$$

for $k\leqq l-1$. The four corner entries of $J_k^{-1}(z)$ can be explicitly determined with the help of (2.2) and (2.11). Solving (6.16) for $w_0(z)$ and $w_k(z)$ we then obtain

$$f(z)p_{k+1}(z)=q_{k+1}(z)+z^{k+1}w_{k+1}(z), \tag{6.17}$$

$$w_k(z)p_{k+1}(z)=1-z+zw_{k+1}(z)p_k(z), \tag{6.18}$$

by use of $w_{-1}(z)=1-z$ and $w_0(z)=f(z)$. Note that (6.17) can be used to define $q_{k+1}(z)$ and $w_{k+1}(z)$ from $f(z)$ and $p_{k+1}(z)$. Furthermore, note that (6.18) can be deduced from (6.17) and (2.14). Dividing (6.18) by $1-z$, using (6.10) and (6.4), we deduce the identity

$$\mu(f_0)\mu(f_1)\cdots\mu(f_k)[p_{k+1}(1)-\mu(f_{k+1})p_k(1)]=1. \tag{6.19}$$

Alternatively, this can be derived from (6.8) and (2.10).

Assume first the CF problem to admit a solution $f(z)$. In view of Lemma 1 we have $0<\mu(f_0)\leqq\infty$, $0<\mu(f_k)<\infty$ for $k=1,\cdots,l-1$, and $0\leqq\mu(f_l)<\infty$, with $\mu(f_l)=0$ in the degenerate case. Hence (6.19) yields $\lambda_{k+1}\geqq\mu(f_{k+1})\geqq 0$, which implies that $J_{l-1}(1)$ is nonnegative definite. Note incidentally that we have $\mu(f_k)=\lambda_k$ for all k if and only if $\mu(f_0)=\infty$, which means that $f(z)$ has no quasizero at $z=1$.

Conversely, assume $J_{l-1}(1)$ to be nonnegative definite, which implies $\lambda_k>0$ for $1\leqq k\leqq l-1$ and $\lambda_l\geqq 0$. Let us choose any C-function $f_l(z)$ satisfying $0\leqq\mu(f_l)\leqq\lambda_l$. We have to show that the function $f_0(z)$ obtained by repeated use of (6.5) belongs to class C. (Then $f(z)=f_0^{-1}(z)$ is a solution to the CF problem.) In view of Lemma 1 it suffices to prove that the constraint (6.6) is satisfied for $k=l-1,l-2,\cdots,0$. To that end let us make use of the descending recurrence relation

$$\mu^{-1}(f_k)-\lambda_k^{-1}=\lambda_{k+1}-\mu(f_{k+1}), \tag{6.20}$$

deduced from (6.19), or directly from (6.8) and (2.10). By induction, (6.20) yields $0\leqq\mu(f_k)\leqq\lambda_k$ whence the desired property $0\leqq\mu(f_k)\leqq 2\,\mathrm{Re}\,\alpha_{k-1}$ for $1\leqq k\leqq l$.

It remains to describe the set of solutions. In the nondegenerate case ($l=n+1$) we obtain the expression (6.14), with $g(z)=f_{n+1}(z)$, by elimination of $w_{n+1}(z)$ and $w_n(z)$ between (6.9) and two versions of (6.17). When $J_n(1)$ is singular, the only possibility is $g(z)=\infty$, which yields $f(z)=q_{n+1}(z)/p_{n+1}(z)$. Similarly, in the degenerate case ($l\leqq n$) we must have $f_l(z)=\infty$, whence $f(z)=q_l(z)/p_l(z)$. These rational functions $f(z)$ are clearly lossless. Their degree properties are described in § 5. □

It is very interesting to compare the formulas (6.3) and (6.14), which must be equivalent. Expressing $p_n(z)$, $q_n(z)$ and $p_{n+1}(z)$, $q_{n+1}(z)$ in terms of $a_n(z)$, $r_n(z)$ by means of (3.7) and (3.8) we obtain a simple relation between the "parameter functions" $\phi(z)$ and $g(z)$ occurring in (6.3) and (6.14), namely

$$g(z)=\frac{1}{2\lambda_{n+1}}\left[\frac{1+z}{1-z}+\frac{1+\bar{\varepsilon}_{n+1}\phi(z)}{1-\bar{\varepsilon}_{n+1}\phi(z)}\right]. \tag{6.21}$$

This establishes a bijection between the S-functions $\phi(z)$ and the C-functions $g(z)$ satisfying (6.15), in perfect agreement with Theorem 2.

Using Lemma 1 in the opposite direction we can obtain an interesting new criterion to check whether a certain function $f(z)$, given by its Maclaurin expansion, belongs to class C. This criterion is expressed in terms of the sequence of complex numbers $\alpha_k = f_k(0)$, with $k = 0, 1$, etc., deduced from the recurrence relation (6.5) applied to the initial function $f_0(z) = f^{-1}(z)$. An important distinction has to be made between the *regular case*, where the sequence of α_k is infinite, and the *singular case*, where the sequence has finite length $n+1$ (ending with α_n) because the algorithm yields $f_{n+1}(z) = \infty$.

THEOREM 3. *The given function $f(z)$ belongs to class C if and only if the tridiagonal matrix $J_k(1)$ built from the parameters $\alpha_t = f_t(0)$ is positive definite, for all k, in the regular case, and is nonnegative definite for $k = n$ in the singular case.*

Proof. The "only if" part follows directly from Theorem 2. The "if" part can be deduced from the Carathéodory-Toeplitz theorem with the help of the congruence relation (4.13). □

It should be noted that the singular case corresponds exactly to the case of a rational lossless function $f(z)$, whose degree equals the rank of $J_n(1)$. As an interpretation of Theorem 3 one can say that a C-function $f(z)$ admits a formal continued fraction expansion

$$f(z) = \frac{1|}{|u_0(z)} + \frac{1|}{|u_1(z)} + \frac{1|}{|u_2(z)} + \cdots, \tag{6.22}$$

with $u_t(z) = (\alpha_t + \bar{\alpha}_t z)/(1-z)$ for even t and $u_t(z) = (1-z^{-1})(\alpha_t + \bar{\alpha}_t z)$ for odd t, yielding nonnegative definite matrices $J_k(1)$. This fraction contains a finite number of terms if and only if $f(z)$ is a rational lossless function. More precisely, if (6.22) contains $n+1$ terms then $f(z)$ has degree n or $n+1$ according as $J_n(1)$ is singular or not.

To conclude this section let us explain how the basic relations (6.12) and (6.13) of our approach to the CF problem give rise to a new efficient "Schur-type algorithm" computing the reflection coefficients ρ_k for a given nonnegative definite Toeplitz matrix C_n (of rank n or $n+1$). In the case of real data, the method described below reduces to the split Schur algorithm recently proposed by the authors [6]. The formulas (6.12) and (6.13) allow us to determine the sequence of parameters $\alpha_0, \alpha_1, \cdots, \alpha_n$ from the data $w_{0,0} = \gamma_0$ (with $\operatorname{Re}\gamma_0 = c_0$) and $w_{0,t} = 2c_t$ for $1 \leqq t \leqq n$. Then the reflection coefficients $\rho_1, \rho_2, \cdots, \rho_n$ can be recursively computed with the help of (3.13). In explicit form, we have

$$\rho_k = \varepsilon_k - [\alpha_k \alpha_{k-1}(\bar{\varepsilon}_k + \bar{\rho}_{k-1})]^{-1}, \tag{6.23}$$

for $k = 1, 2, \cdots, n$, with the initialization $\rho_0 = 1$. The number ε_k appearing in (6.23) can be determined by

$$\varepsilon_k = w_{k-1,0} / \bar{w}_{k-1,0}. \tag{6.24}$$

Indeed, we have $w_k(0) p_{k+1}(0) = 1$ by (6.18), so that (6.24) results from the definition (3.9). It can be verified that the number of multiplications required by this method (based on (6.12), (6.13), (6.23), and (6.24)) is approximately reduced by a factor 2 with respect to the classical Schur algorithm, while the number of additions remains unchanged. These conclusions are very similar to those concerning the Levinson-type algorithm of § 3.

7. Nevanlinna–Pick interpolation problem. Given $n+1$ distinct points $z_0, z_1, \cdots, z_n$ in the unit disk $|z| < 1$ and $n+1$ complex numbers $u_0, u_1, \cdots, u_n$, the

Nevanlinna-Pick (NP) *interpolation problem* requires to determine the set of C-functions $f(z)$ satisfying

$$f(z_k) = u_k \quad \text{for } k = 0, 1, \cdots, n. \tag{7.1}$$

This appears as a natural analogue of the CF problem. It is classically solved by means of the Nevanlinna algorithm [2], which is an extension of the Schur algorithm. Let us now explain how the NP problem can alternatively be approached by a suitable modification of the "tradiagonal method" used in § 6. From the given point z_k define the binomial

$$y_k(z) = \beta_k \frac{z - z_k}{1 - z_k} \quad \text{with } \beta_k = \frac{|1 - z_k|^2}{1 - |z_k|^2}. \tag{7.2}$$

The Blaschke function $y_k(z)/\hat{y}_k(z)$ acts as a transformation that preserves the unit disk, fixes the point $z = 1$ and maps the point $z = z_k$ to the origin $z = 0$. Therefore, the following key result is obtained from Lemma 1 by a simple change of variables.

LEMMA 4. *Let $f_k(z)$ be a C-function and define $f_{k+1}(z)$ by means of the relation*

$$f_k(z) = \frac{\alpha_k \hat{y}_k(z) + \bar{\alpha}_k y_k(z)}{1 - z} - \frac{y_k(z)\hat{y}_k(z)}{(1 - z)^2 f_{k+1}(z)}, \tag{7.3}$$

where $\alpha_k = f_k(z_k)$ and $\hat{y}_k(z) = z\bar{y}_k(1/\bar{z})$. Then $f_{k+1}(z)$ is a C-function satisfying

$$0 \leqq \mu(f_{k+1}) \leqq 2\beta_k^{-1} \operatorname{Re} \alpha_k, \tag{7.4}$$

with the possibility $f_{k+1}(z) = \infty$. Conversely, let α_k be a complex number and let $f_{k+1}(z)$ be a C-function, subject to the constraint (7.4). *Then the right-hand side of* (7.3) *is a C-function satisfying $f_k(z_k) = \alpha_k$.*

It is easily seen that Lemma 4 yields an iterative solution method for the NP problem, like Lemma 1 for the CF problem. Let us briefly describe the method in question. Define the functions $w_k(z)$ exactly as in (6.9), (6.10). Then (7.3) can be written in the form of the recurrence relation

$$y_k(z)\hat{y}_k(z)w_{k+1}(z) = (\alpha_k \hat{y}_k(z) + \bar{\alpha}_k y_k(z))w_k(z) - w_{k-1}(z), \tag{7.5}$$

for $k = 0, 1, \cdots, n$. The initial conditions are given by $w_{-1}(z) = 1 - z$ and $w_0(z) = f(z)$, yielding $f_0(z) = f^{-1}(z)$, where $f(z)$ is supposed to be a solution of the NP problem (7.1).

Thus, except in some "pathological situations," the interpolation constraints (7.1) determine the $n+1$ parameters $\alpha_0, \alpha_1, \cdots, \alpha_n$ via the identities

$$\alpha_k = w_{k-1}(z_k)/(1 - z_k)w_k(z_k), \tag{7.6}$$

$$w_{k+1}(z_j) = [(\alpha_k \hat{y}_k(z_j) + \bar{\alpha}_k y_k(z_j))w_k(z_j) - w_{k-1}(z_j)]/y_k(z_j)\hat{y}_k(z_j). \tag{7.7}$$

In principle, they have to be used for $0 \leqq k \leqq n$ and $k+1 \leqq j \leqq n$, with the initial conditions $w_{-1}(z_j) = 1 - z_j$ and $w_0(z_j) = u_j$. But the algorithm has to stop if it meets the situation $w_l(z_l) = 0$. Then it is clear that the NP problem can have a solution only if $w_l(z_j) = 0$ for $j = l, l+1, \cdots, n$; this will be called the *degenerate case*. The normal situation where $w_k(z_k) \neq 0$ for all k will be called the *nondegenerate case* and is conveniently characterized by $l = n+1$.

As explained below, the complete solution of the NP problem can be expressed in terms of an appropriate tridiagonal matrix $J_k(z)$ built from the data z_j and α_j. The

definition is

$$J_k(z)=\begin{bmatrix} \alpha_0\hat{y}_0(z)+\bar{\alpha}_0 y_0(z) & y_0(z) & & \\ \hat{y}_0(z) & & \ddots & \\ & \ddots & & y_{k-1}(z) \\ & & \hat{y}_{k-1}(z) & \alpha_k\hat{y}_k(z)+\bar{\alpha}_k y_k(z) \end{bmatrix}, \tag{7.8}$$

for $k=0,1,\cdots,l-1$. Note that $y_k(1)=\hat{y}_k(1)=\beta_k$; hence $J_k(1)$ is a real symmetric matrix. From $J_k(z)$ define the polynomials $p_k(z)=\det J_{k-1}(z)$ and $q_k(z)=(1-z)\det J^0_{k-1}(z)$, exactly as in (2.2) and (2.11). They enjoy the symmetry and antisymmetry properties (2.3) and (2.13). Furthermore, they obey the recurrence relations

$$p_{k+1}(z)=(\alpha_k\hat{y}_k(z)+\bar{\alpha}_k y_k(z))p_k(z)-y_{k-1}(z)\hat{y}_{k-1}(z)p_{k-1}(z), \tag{7.9}$$

$$q_{k+1}(z)=(\alpha_k\hat{y}_k(z)+\bar{\alpha}_k y_k(z))q_k(z)-y_{k-1}(z)\hat{y}_{k-1}(z)q_{k-1}(z), \tag{7.10}$$

with the initial conditions $p_{-1}(z)=0$, $p_0(z)=1$ and $q_0(z)=0$, $q_1(z)=1-z$. As a generalization of (2.14), we have the identity

$$p_k(z)q_{k+1}(z)-q_k(z)p_{k+1}(z)=(1-z)\prod_{i=0}^{k-1} y_i(z)\hat{y}_i(z). \tag{7.11}$$

Define the *Jacobi parameters* $\lambda_k=p_k(1)/p_{k-1}(1)$ as in (2.8). It is clear that the matrix $J_n(1)$ is nonnegative definite if and only if the positivity constraints (2.9) are satisfied. (In this case, $J_{n-1}(1)$ is positive definite.) The Jacobi parameters can be recursively computed from the data α_k and β_k by the formula

$$\lambda_{k+1}=2\beta_k \operatorname{Re}\alpha_k-\beta_{k-1}^2\lambda_k^{-1}, \tag{7.12}$$

which directly follows from (7.9). It is a generalization of (2.10). The initialization is given by $\lambda_0=\infty$.

The solution of the Nevanlinna-Pick interpolation problem can be expressed in terms of these ingredients in essentially the same way as in Theorem 2. The only difference lies in the expression of the general solution in the nondegenerate case; it is given by

$$f(z)=\frac{y_n(z)\hat{y}_n(z)q_n(z)-(1-z)g(z)q_{n+1}(z)}{y_n(z)\hat{y}_n(z)p_n(z)-(1-z)g(z)p_{n+1}(z)}, \tag{7.13}$$

where $g(z)$ is a C-function subject to the only constraint

$$0\leqq\mu(g)\leqq\beta_n^{-2}\lambda_{n+1}. \tag{7.14}$$

In particular, if $J_n(1)$ is singular then (7.14) forces $g(z)=\infty$; this yields the unique solution $f(z)=q_{n+1}(z)/p_{n+1}(z)$, which is lossless and has degree n.

The proof will not be repeated here, since it is basically the same as that of Theorem 2. Let us only indicate the appropriate substitutes of the identities (6.8) and (6.17), which are required to derive (7.13). From (7.3) we immediately deduce

$$\beta_k^2\mu(f_{k+1})=2\beta_k \operatorname{Re}\alpha_k-\mu^{-1}(f_k). \tag{7.15}$$

By comparison with (7.12) this allows us to show that the property $\beta_k^2\mu(f_{k+1})\leqq\lambda_{k+1}$ holds for all k if and only if it holds for $k=n$ (under the positivity assumption). Hence, if the condition (7.14) is satisfied, with $g(z)=f_{n+1}(z)$, then so are all conditions (7.4). The representation (7.13) can be obtained with the help of the identity

$$f(z)p_{k+1}(z)=q_{k+1}(z)+\left(\prod_{i=0}^{k} y_i(z)\hat{y}_i(z)\right)w_{k+1}(z), \tag{7.16}$$

which results from a similar computation as in § 6. In fact, we deduce formula (7.13), with $g(z)=f_{n+1}(z)$, by equating the ratio $w_n(z)/w_{n+1}(z)$ derived from (7.16) to the function $(1-z)g(z)$, in agreement with (6.9). It is interesting to mention the analogue of (6.18), which is

$$w_k(z)p_{k+1}(z)=1-z+y_k(z)\hat{y}_k(z)w_{k+1}(z)p_k(z). \tag{7.17}$$

As an illustration of the theory let us finally indicate a generalization of the Bistritz stability test. Consider a complex polynomial $x_m(z)$ of degree m, having the property that $x_m(1)$ is real. Construct both symmetric polynomials

$$b_m(z)=x_m(z)+\hat{x}_m(z), \qquad b_{m-1}(z)=\frac{x_m(z)-\hat{x}_m(z)}{1-z}. \tag{7.18}$$

Choosing m distinct points $z_0, z_1, \cdots, z_{m-1}$ in the unit disk, define complex numbers α_k and symmetric polynomials $b_{m-k-2}(z)$, for $k=0, 1, \cdots, m-1$, from the relations

$$\alpha_k=b_{m-k}(z_k)/(1-z_k)b_{m-k-1}(z_k), \tag{7.19}$$

$$y_k(z)\hat{y}_k(z)b_{m-k-2}(z)=(\alpha_k\hat{y}_k(z)+\bar{\alpha}_k y_k(z))b_{m-k-1}(z)-b_{m-k}(z), \tag{7.20}$$

with the initialization (7.18). By construction, $b_t(z)$ is a polynomial of formal degree t (for $t\geqq 0$), and $b_{-1}(z)$ vanishes. It is interesting to note that $b_{m-k}(z)$ can be written in the form

$$b_{m-k}(z)=b_0 \det J_{m-1}^{k-1}(z), \tag{7.21}$$

for a real constant b_0, where $J_{m-1}^{k-1}(z)$ is the submatrix of $J_{m-1}(z)$ obtained by suppression of its first k rows and columns. The Bistritz stability criterion can be generalized as follows.

THEOREM 5. *The polynomial $x_m(z)$ is stable (i.e., devoid of zeros in the closed unit disk $|z|\leqq 1$) if and only if the tridiagonal matrix $J_{m-1}(1)$ built from the data z_k and α_k exists and is positive definite.*

Proof. Set $f(z)=(1-z)b_{m-1}(z)/b_m(z)$. It is well known that $x_m(z)$ is stable if and only if $f(z)$ is a lossless function of exact degree m. In this case, $f(z)$ can be viewed as a solution to the NP problem (7.1), with $u_k=(1-z_k)b_{m-1}(z_k)/b_m(z_k)$ and $n=m-1$. Let us construct rational functions $f_k(z)$ and complex numbers $\alpha_k=f_k(z_k)$, for $k=0, 1, \cdots, m-1$, by use of the recurrence relation (7.3) with the initialization $f_0(z)=f^{-1}(z)$. By comparison between (7.5) and (7.20) it is seen that the functions (6.10) are related to the polynomials $b_t(z)$ by the simple identity

$$w_k(z)=(1-z)b_{m-k-1}(z)/b_m(z), \tag{7.22}$$

for $-1\leqq k\leqq m$. The desired result is obtained by straightforward application of the generalized version of Theorem 2; the present situation corresponds to the nondegenerate case and the choice $g(z)=\infty$ in (7.13), yielding $f(z)=q_m(z)/p_m(z)$. □

When comparing the method above to that indicated in § 5 (in the "confluent case" $z_j=0$ for all j) one should note that the parameters α_k are numbered in the reverse order, which is a simple matter of notation. Of course, the positive definiteness of $J_{m-1}(1)$ amounts to the property $\lambda_k>0$ for $k=1, 2, \cdots, m$, where the parameters λ_k are determined by use of the recurrence relation (7.12), with $\lambda_0=\infty$. An equivalent version of the criterion of Theorem 5 says that $x_m(z)$ is stable if and only if the values assumed by the polynomials $b_t(z)$ at the point $z=1$ are nonzero and have constant sign for $0\leqq t\leqq m$. This is a straightforward consequence of the identity (7.21).

REFERENCES

[1] A. C. AITKEN, *Determinants and Matrices*, Oliver and Boyd, London, 1958.

[2] N. I. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, London, 1965.

[3] Y. BISTRITZ, *A stability new test for linear discrete systems in a table form*, IEEE Trans. Circuits and Systems, CAS-30 (1983), pp. 917-919.

[4] ———, *Zero location with respect to the unit circle of discrete-time linear system polynomials*, Proc. IEEE, 72 (1984), pp. 1131-1142.

[5] P. DELSARTE AND Y. GENIN, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., ASSP-34 (1986), pp. 470-478.

[6] ———, *On the splitting of classical algorithms in linear prediction theory*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 645-653.

[7] P. DELSARTE, Y. GENIN, AND Y. KAMP, *On the role of the Nevanlinna-Pick problem in circuit and system theory*, Circuit Theory Appl., 9 (1981), pp. 177-187.

[8] ———, *Application of the index theory of pseudo-lossless functions to the Bistritz stability test*, Philips J. Res., 39 (1984), pp. 226-241.

[9] ———, *Pseudo-Carathéodory functions and Hermitian Toeplitz matrices*, Philips J. Res., 41 (1986), pp. 1-54.

[10] P. DELSARTE, Y. GENIN, Y. KAMP, AND P. VAN DOOREN, *Speech modeling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277-292.

[11] T. T. GEORGIOU, *Realization of power spectra from partial covariance sequences*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 438-449.

[12] L. YA. GERONIMUS, *Orthogonal Polynomials*, Consultants Bureau, New York, 1961.

[13] J. GILEWICZ AND E. LEOPOLD, *Location of the zeros of polynomials satisfying three-term recurrence relations. I. General case with complex coefficients*, J. Approx. Theory, 43 (1985), pp. 1-14.

[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford Academic, Oxford, 1983.

[15] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and their Applications*, University of California Press, Berkeley, CA, 1958.

[16] M. H. GUTKNECHT, J. O. SMITH, AND L. N. TREFETHEN, *The Carathéodory-Fejér method for recursive digital filter design*, IEEE Trans. Acoust. Speech Signal Process., ASSP-31 (1983), pp. 1417-1426.

[17] J. W. HELTON, *The distance of a function to H^∞ in the Poincaré metric; electrical power transfer*, J. Funct. Anal., 38 (1980), pp. 273-314.

[18] P. HENRICI, *Applied and Computational Complex Analysis*, Vol. I, John Wiley, New York, 1974.

[19] E. I. JURY, *Theory and Application of the z-Transform Method*, John Wiley, New York, 1964.

[20] T. KAILATH, *A view of three decades of linear filtering theory*, IEEE Trans. Inform. Theory, IT-20 (1974), pp. 145-181.

[21] N. LEVINSON, *The Wiener rms (root mean square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1946), pp. 261-278.

[22] J. MAKHOUL, *Linear prediction: a tutorial review*, Proc. IEEE, 63 (1975), pp. 561-580.

[23] M. MARDEN, *Geometry of Polynomials*, American Mathematical Society, Providence, RI, 1966.

[24] L. R. RABINER AND R. W. SCHAFER, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

[25] D. SARASON, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179-203.

[26] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, New York, 1959.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

016

ON COMPLETE SYMMETRIC FUNCTIONS*

EDWARD NEUMAN†

Abstract. This paper is devoted to the study of some properties of the complete symmetric functions. These functions play an important role in the theory of partitions and in the combinatorics as well. Among other things, the representation formulas as well as the recurrence formulas and inequalities involving functions under discussion are given. Some applications are also included.

Key words. complete symmetric functions, divided differences, B-splines, q-binomial coefficients, r-Stirling numbers of the second kind

AMS(MOS) subject classification. primary 26C05

1. Introduction. Let $(x_0, \cdots, x_n) \in \mathbb{R}^{n+1}$ $(n \geqq 0)$. The rth complete symmetric function h_r $(r = 0, 1, \cdots)$ in the variables $x_0, \cdots, x_n$ is defined by

$$h_r \equiv h_r(x_0, \cdots, x_n) = \sum_{i_0+\cdots+i_n=r} x_0^{i_0} \cdots x_n^{i_n}, \tag{1.1}$$

where $i_0, \cdots, i_n \in \{0, 1, \cdots, r\}$ (see, e.g., [6]). The sum (1.1) involves $\binom{n+r}{r}$ terms. Without loss of generality we may assume $x_0 \leqq x_1 \leqq \cdots \leqq x_n$. Setting $x_i = q$ $(0 \leqq i \leqq n)$, we obtain $h_r = \binom{n+r}{r} q^r$. In what follows we will assume $x_0 < x_n$.

The functions h_r play an important role in the theory of partitions (see [1] for more details). Also they are useful in combinatorics. Letting $x_i = q^i$ $(0 \leqq i \leqq n)$, where q is an indeterminate, we obtain

$$h_r = \begin{bmatrix} n+r \\ r \end{bmatrix},$$

where as usual

$$\begin{bmatrix} m \\ k \end{bmatrix} = \frac{(q^m-1)(q^m-q) \cdot \cdots \cdot (q^m-q^{k-1})}{(q^k-1)(q^k-q) \cdot \cdots \cdot (q^k-q^{k-1})}$$

denotes the q-binomial coefficient or Gaussian polynomial (see, e.g., [1]). Another choice for x_i, namely $x_i = r+i$ $(0 \leqq i \leqq n;\ r = 0, 1, \cdots)$ leads to

$$h_m(r, r+1, \cdots, r+n) = S_r(m+n+r, n+r) \qquad (m, n, r \in \mathbb{N}_0) \tag{1.2}$$

(see § 7 for the proof of (1.2)). Here $S_r(\cdot, \cdot)$ stands for the r-Stirling number of the second kind. The number $S_r(k, n)$ $(0 \leqq r \leqq n \leqq k)$ is defined combinatorially as the number of partitions of the set $\{1, 2, \cdots, k\}$ into n nonempty disjoint subsets, such that the numbers $1, 2, \cdots, r$ are in distinct subsets (cf. [3]). The classical Stirling numbers of the second kind $S(k, n)$ coincide with $S_r(k, n)$ when $r = 0$ or $r = 1$.

In our proofs the B-splines of Curry and Schoenberg [5] play an important role. Some elementary properties of these functions are given in § 2. The representation formulas for h_r are discussed in § 3. In § 4 we deal with the generating functions for h_r. Some new recurrence formulas for h_r are given in § 5. Inequalities involving the functions under discussion are given in § 6. Section 7 is devoted to applications.

2. Preliminaries. Throughout this paper we will denote by $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}_0$, and $\mathbb{N}$ the sets of all reals, integers, nonnegative integers and positive integers, respectively.

* Received by the editors September 2, 1986; accepted for publication June 4, 1987.
† Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901.

For our further aims we introduce the B-splines of Curry and Schoenberg [5]. Let $\cdots \leqq t_{-1} \leqq t_0 \leqq t_1 \leqq \cdots$ be a bi-infinite partition of $\mathbb{R}$ with at most n ($n \in \mathbb{N}$) values of the t's equal to each other, i.e., $t_i < t_{i+n}$ for all $i \in \mathbb{Z}$. The function

$$M_{i,n}(t) = n[t_i, \cdots, t_{i+n}](\cdot - t)_+^{n-1}$$

is the B-spline of degree $n-1$ (order n) with knots at $t_i, \cdots, t_{i+n}$. As usual $[t_i, \cdots, t_{i+n}]g$ denotes the divided difference of order n for the function g at the points t_{i+l} ($0 \leqq l \leqq n$) and

$$(x-t)_+^{n-1} = (\text{Max}\,\{0, x-t\})^{n-1}$$

denotes the truncation power function. For the reader's convenience we list below some well-known properties of the B-splines:

(i) $M_{i,n}(t) > 0$ for $t \in (t_1, t_{i+n})$ and $M_{i,n}(t) = 0$ otherwise. Thus supp $M_{i,n} = [t_i, t_{i+n}]$.

(ii) In each interval $[t_{i+j}, t_{i+j+1}]$ ($t_{i+j} < t_{i+j+1}$; $j = 0, 1, \cdots, n-1$) $M_{i,n}$ coincides with an algebraic polynomial of degree $n-1$ or less.

(iii) Let t_{i+j} be a knot of multiplicity k, i.e., let $t_{i+j-1} < t_{i+j} = \cdots = t_{i+j+k-1} < t_{i+j+k}$, then $M_{i,n}$ is exactly $n-1-k$ times continuously differentiable on (t_{i+j-1}, t_{i+j+k}).

(iv) If $f \in C^n[t_i, t_{i+n}]$, then

$$[t_i, \cdots, t_{i+n}]f = \frac{1}{n!} \int_{t_i}^{t_{i+n}} M_{i,n}(t) f^{(n)}(t)\, dt.$$

(v) The following recurrence formula

$$\frac{n-1}{n} M_{i,n}(t) = \frac{t-t_i}{t_{i+n}-t_i} M_{i,n-1}(t) + \frac{t_{i+n}-t}{t_{i+n}-t_i} M_{i+1,n-1}(t)$$

$$(i \in \mathbb{Z},\ n \geqq 2,\ t \in \mathbb{R})$$

holds true (see, e.g., [14]).

3. Representation formulas for the complete symmetric functions. The formula (1.1) seems to be rather inconvenient for our further purposes. In this section we offer several equivalent formulas for h_r. First of all we need more notation. Let $x_0 \leqq x_1 \leqq \cdots \leqq x_n$. In order to describe exactly where the equalities hold, we suppose that

$$x_0 \leqq x_1 \leqq \cdots \leqq x_n = \underbrace{\tau_0, \cdots, \tau_0}_{l_0}, \cdots, \underbrace{\tau_d, \cdots, \tau_d}_{l_d}, \tag{3.1}$$

where each τ_i is repeated exactly l_i times with $\sum_{i=0}^{d} l_i = n+1$. Then given any sufficiently differentiable functions $u_0, \cdots, u_n$, we define a matrix

$$\begin{bmatrix} x_0, \cdots, x_n \\ u_0, \cdots, u_n \end{bmatrix} := [D^{d_i} u_j(x_i)]_{i,j=0}^{n} \tag{3.2}$$

with $d_i = \text{Max}\,\{j\colon x_i = \cdots = x_{i+j}\}$, $i = 0, 1, \cdots, n$, where, as usual, D^m denotes the operator of differentiation. Further let

$$S^n = \left\{ (\lambda_1, \cdots, \lambda_n) \in \mathbb{R}^n \colon \lambda_i \geqq 0, \text{ for all } i, \sum_{i=1}^{n} \lambda_i \leqq 1 \right\}$$

denote the n-simplex.

Also let C denote a simple closed contour enclosing a simply connected region of the complex variable z in which are situated the points $\tau_0, \cdots, \tau_d$.

We are ready to state and prove the main result of this section.

THEOREM 3.1. *Let* $r \in \mathbb{N}_0$. *Then*

$$h_r = [x_0, \cdots, x_n]t^{n+r}, \tag{3.3}$$

and also

$$h_r = \binom{n+r}{r} \int_{x_0}^{x_n} M_{0,n}(t)t^r \, dt, \tag{3.4}$$

where $M_{0,n}$ *denotes the B-spline of order n with knots at* $x_0, \cdots, x_n$. *Also*

$$h_r = (n+r) \cdots (r+1) \int_{S^n} \left[\sum_{i=0}^{n} \lambda_i x_i \right]^r d\lambda, \tag{3.5}$$

where $\lambda_0 = 1 - \lambda_1 - \cdots - \lambda_n$ *and* $d\lambda = d\lambda_1 \cdots d\lambda_n$. *Moreover, if* τ_i *and* l_i *are the same as in* (3.1), *then*

$$h_r = \frac{1}{2\pi i} \int_C \frac{z^{n+r} \, dz}{(z-\tau_0)^{l_0} \cdots (z-\tau_d)^{l_d}}, \tag{3.6}$$

and also the following formula:

$$h_r = \det \begin{bmatrix} x_0 & x_1, \cdots, x_{n-1}, x_n \\ 1 & t, \cdots, t^{n-1}, t^{n+r} \end{bmatrix} \Big/ \det \begin{bmatrix} x_0, \cdots, x_n \\ 1, \cdots, t^n \end{bmatrix} \tag{3.7}$$

holds true.

Proof. The formula (3.3) has been established in [8]. In order to prove (3.4) we set $t_{i+j} = x_j$ $(j = 0, 1, \cdots, n)$, next $i = 0$ and $f(t) = t^{n+r}$ into (iv). Hence and from (3.3) the desired result follows. For the proof of (3.5) we apply the Hermite-Genocchi formula for divided differences

$$[x_0, \cdots, x_n]f = \int_{S^n} f^{(n)} \left[\sum_{i=0}^{n} \lambda_i x_i \right] d\lambda \qquad (f \in \mathbb{C}^n[x_0, x_n]) \tag{3.8}$$

(cf. [2]). Setting above $f(t) = t^{n+r}$, then making use of (3.3), we obtain (3.5). Direct applications of (3) in [8, p. 14] to (3.3) gives the assertion (3.6). Equation (3.7) follows immediately from (2.86) in [14] and from (3.3). This completes the proof. □

COROLLARY 3.1 [13]. *Let* $x_j = -\cos(\pi j/n)$ $(j = 0, 1, \cdots, n)$ *be the Chebyshev points. Then*

$$h_r = \begin{cases} 2^{-2m} \sum_{i=0}^{[m/n]} \binom{n+2m}{m-in} & \text{if } r = 2m \\ 0 & \text{if } r = 2m+1 \end{cases} \qquad (m = 0, 1, \cdots), \tag{3.9}$$

where, as usual, $[x]$ *denotes the largest integer not bigger than x.*

Proof. First we will establish the first formula of (3.9). Let g be an integrable function on $[-1, 1]$ and let

$$a_j[g] = \frac{2}{\pi} \int_{-1}^{1} \frac{g(t)T_j(t)}{\sqrt{1-t^2}} \, dt$$

denote the jth Fourier-Chebyshev coefficient of g with T_j—the jth Chebyshev polynomial of the first kind. We will use the following result due to S. Bernstein:

$$[x_0, \cdots, x_n]g = 2^{n-1} \sum_{i=0}^{\infty} a_{(2i+1)n}[g] \tag{3.10}$$

(see, e.g., [12]). Also it is known that

$$t^{n+2m} = 2^{-n-2m+1} \sum_{j=0}^{[n/2]+m\,\circ} \binom{n+2m}{j} T_{n+2m-2j}(t) \tag{3.11}$$

where

$$\sum_{j=l}^{m\,\circ} a_j T_{k_j} := \sum_{j=l}^{m} \left[\frac{1}{2} a_j \text{ if } k_j = 0,\ a_j \text{ if } k_j \neq 0\right] T_{k_j}$$

(cf. [12, (2.40)]). Hence we obtain

$$a_{(2i+1)n}[t^{n+2m}] = 2^{-n-2m+1}\binom{n+2m}{m-in}. \tag{3.12}$$

Combining (3.3), (3.10), and (3.12) yields

$$h_{2m} = [x_0, \cdots, x_n]t^{n+2m} = 2^{-2m} \sum_{i=0}^{[m/n]} \binom{n+2m}{m-in}.$$

The second formula of (3.9) follows immediately from (3.4) and from the fact that the B-spline $M_{0,n}$ is an even function in the case under consideration. The proof is completed. □

COROLLARY 3.2 [13]. *Let* $x_j = -\cos \alpha_j$, *where* $\alpha_j = (2j+1)\pi/(2n+2)$ $(j = 0, 1, \cdots, n)$ *be the zeros of* T_{n+1}. *Then*

$$h_r = \begin{cases} 2^{-2m}(-1)^n \left\{ \sum_{q=0}^{[m/(n+1)]} \binom{n+2m}{m-q(n+1)} \right. \\ \qquad \left. - \sum_{q=0}^{[(m-1)/(n+1)]} \binom{n+2m}{m-1-q(n+1)} \right\} & \text{if } r = 2m, \\ 0 \quad \text{if } r = 2m+1 \end{cases} \tag{3.13}$$

$(m = 0, 1, \cdots,).$

Proof. It is a well-known fact that

$$[x_0, \cdots, x_n]g = (-1)^{n+1} \frac{2^n}{n+1} \sum_{j=0}^{n} (-1)^j \sin \alpha_j g(x_j).$$

In order to prove the first formula of (3.13) we set above $g(t) = t^{n+2m}$. Hence and by virtue of (3.3) we get

$$h_{2m} = [x_0, \cdots, x_n]t^{n+2m} = \frac{-2^n}{n+1} \sum_{j=0}^{n} (-1)^j \sin \alpha_j \cos^{n+2m} \alpha_j.$$

The last sum depends upon the numbers $\mathscr{D}_{nk}$, where

$$\mathscr{D}_{nk} := \sum_{j=0}^{n} (-)^j \sin \alpha_j \cos kj \quad (k\text{-integer}).$$

It is easy to verify that

$$\mathscr{D}_{nk} = \frac{1}{2}(E_{n,k+1} - E_{n,k-1}), \tag{3.14}$$

where

$$E_{nk} := \sum_{j=0}^{n} (-1)^j \sin k\alpha_j.$$

Hence, we obtain

$$\left[\cos\frac{k\pi}{2n+2}\right]E_{nk}=\sum_{j=0}^{n}(-1)^j\sin\frac{(2j+1)k\pi}{2n+2}\cos\frac{k\pi}{2n+2}$$
$$=\frac{1}{2}\sum_{j=0}^{n}(-1)^j\left\{\sin\frac{(j+1)k\pi}{n+1}+\sin\frac{jk\pi}{n+1}\right\}$$
$$=\frac{1}{2}(-1)^n\sin\frac{(n+1)k\pi}{n+1}=0.$$

Therefore if $k\pi/(2n+2)\neq(p+\frac{1}{2})\pi$ (p-integer), i.e., if $k\neq(2p+1)(n+1)$, then $E_{nk}=0$. Assume now $k=(2p+1)(n+1)$. Then

$$E_{nk}=\sum_{j=0}^{n}(-1)^j\sin\frac{(2j+1)(2p+1)(n+1)\pi}{2n+2}=(-1)^p(n+1).$$

Hence and from (3.14) we conclude that

$$\mathscr{D}_{nk}=\begin{cases}\frac{1}{2}(-1)^{k+1}(n+1) & \text{if } \frac{k+1}{n+1}\text{ is odd},\\ -\frac{1}{2}(-1)^{k-1}(n+1) & \text{if } \frac{k-1}{n+1}\text{ is odd},\\ 0 & \text{otherwise.}\end{cases}\tag{3.15}$$

Setting $t=\cos\rho$ in (3.11), we obtain

$$\cos^{n+2m}\alpha_j=2^{-n-2m+1}\sum_{i=0}^{[n/2]+m}{}^{\circ}\binom{n+2m}{i}\cos(n+2m-2i)\alpha_j.$$

Therefore

$$h_{2m}=\frac{-2^n}{n+1}2^{-n-2m+1}\sum_{i=0}^{[n/2]+m}{}^{\circ}\binom{n+2m}{i}\mathscr{D}_{n,n+2m-2i}.\tag{3.16}$$

In the last sum the only nonzero terms are those for which

$$\frac{n+2m-2i+1}{n+1}\text{ is odd}\quad\text{or}\quad\frac{n+2m-2i-1}{n+1}\text{ is odd},$$

i.e., if $m-i=q(n+1)$ (q-integer) or $m-i-1=q(n+1)$, respectively. Hence one gets $i=m-q(n+1)$ or $i=m-1-q(n+1)$. Simultaneously $0\leqq i\leqq[n/2]+m$. Therefore $0\leqq q\leqq[m/(n+1)]$ or $0\leqq q\leqq[(m-1)/(n+1)]$. Hence and from (3.16) and (3.15) the assertion follows. For the proof that $h_r=0$ if r is odd we apply the same arguments like those in the proof of Corollary 3.1. This completes the proof. □

The functions h_r are well defined when $r\in\mathbb{N}_0$. In many places they are defined to be zero when $r=-1,-2,\cdots$ (see, e.g., [6]). Assuming that (3.3)–(3.7) are the defining formulas for h_r with $r\in\mathbb{Z}$ we get $h_r=0$ for $r=-1,-2,\cdots,-n$. If $r=-n-1,-n-2,\cdots$, then h_r is not necessarily equal to zero. This confirms the following.

COROLLARY 3.3. *Let $x_j\neq 0$ for all j. Then for $l=1,2,\cdots$, we have*

$$h_{-n-l}=\frac{(-1)^n}{\prod_{j=0}^{n}x_j^l}\det\begin{bmatrix}[x_0,x_1]t^l & [x_1]t^l & \cdots & 0\\ \vdots & \vdots & \cdots & \vdots\\ [x_0,\cdots,x_{n-1}]t^l & [x_1,\cdots,x_{n-1}]t^l & \cdots & [x_{n-1}]t^l\\ [x_0,\cdots,x_n]t^l & [x_1,\cdots,x_n]t^l & \cdots & [x_{n-1},x_n]t^l\end{bmatrix}.\tag{3.17}$$

Hence in particular

$$h_{-n-1}=(-1)^n\Big/\prod_{j=0}^{n}x_j \tag{3.18}$$

and

$$h_{-n-2}=(-1)^n\left\{\sum_{j=0}^{n}\prod_{\substack{l=0\\ l\neq j}}^{n}x_l\right\}\Big/\prod_{j=0}^{n}x_j^2. \tag{3.19}$$

Proof. In order to prove the identity (3.17) we apply the following formula:

$$[x_0,\cdots,x_n]\frac{1}{f(x)}=\frac{(-1)^n}{\prod_{j=0}^{n}f(x_j)}\det\begin{bmatrix}[x_0,x_1]f & [x_1]f & 0 & \cdots & 0\\ \vdots & \vdots & & \cdots & \vdots\\ [x_0,\cdots,x_n]f & [x_1,\cdots,x_n]f & & \cdots & [x_{n-1},x_n]f\end{bmatrix}$$

which holds true provided $f(x_j)\neq 0$ for all j (see [11, Ex. 2.4.19]). Setting above $f(t)=t^l$ and next making use of (3.3), we arrive at (3.17). The identities (3.18) and (3.19) follow immediately from (3.17). The proof is completed. □

Combining (3.5) and (3.18) yields

$$\int_{S^n}\left[\sum_{i=0}^{n}\lambda_i x_i\right]^{-n-1}d\lambda=1/n!\prod_{j=0}^{n}x_j\qquad(x_j\neq 0).$$

The last identity is due to R. Feynman.

We close the section with a differentiation formula for h_r.

COROLLARY 3.4. *Let x_j be of the multiplicity l_j $(0\leqq j\leqq n;\ 1\leqq l_j\leqq n)$. Then*

$$\frac{\partial^l}{\partial x_j^l}h_r(x_0,\cdots,x_{j-1},\underbrace{x_j,\cdots,x_j}_{l_j},x_{j+1},\cdots,x_n)$$

$$=l_j(l_j+1)\cdots(l_j+l-1)h_r(x_0,\cdots,x_{j-1},\underbrace{x_j,\cdots,x_j}_{l_j+l},x_{j+1},\cdots,x_n).$$

Proof. The last result follows immediately from (3.3) and from the well-known formula for the divided differences (see [11]). □

4. Generating functions. Let H denote a generating function for the complete symmetric functions, i.e., let

$$H(t)=\sum_{r=0}^{\infty}h_r t^r.$$

Then

$$H(t)=\prod_{i=0}^{n}(1-x_i t)^{-1}$$

(see, e.g., [6]). In this section we give simple formulas for the exponential generating function E, where

$$E(t)=\sum_{r=0}^{\infty}h_r\frac{t^{n+r}}{(n+r)!}\qquad(n\in\mathbb{N}).$$

We have the following.

THEOREM 4.1. *Let* $n \in \mathbb{N}$. *Then*

$$E(t) = [x_0, \cdots, x_n]_{(x)}\, e^{xt}, \tag{4.1}$$

where the subscript (x) *denotes that the divided difference operator acts on the variable* x. *Also*

$$E(t) = \frac{t^n}{n!} \int_{x_0}^{x_n} M_{0,n}(x)\, e^{xt}\, dx, \tag{4.2}$$

where $M_{0,n}$ *denotes the B-spline of order* n *with knots at* $x_0, \cdots, x_n$. *Moreover, if* $x_i \neq x_j$ *for* $i \neq j$, *then*

$$E(t) = \sum_{j=0}^{n} e^{x_j t}/w_j \tag{4.3}$$

with

$$w_j = \prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i) \qquad (0 \leqq j \leqq n).$$

Proof. First we will show that the following identity

$$[x_0, \cdots, x_n]f = \sum_{r=0}^{\infty} h_r \frac{f^{(n+r)}(0)}{(n+r)!} \tag{4.4}$$

holds true provided f is sufficiently smooth. We have

$$f(x) = \sum_{r=0}^{\infty} f^{(r)}(0) \frac{x^r}{r!} = \sum_{r=-n}^{\infty} f^{(n+r)}(0) \frac{x^{n+r}}{(n+r)!}.$$

Hence

$$[x_0, \cdots, x_n]f = \sum_{r=-n}^{\infty} ([x_0, \cdots, x_n]x^{n+r}) \frac{f^{(n+r)}(0)}{(n+r)!}.$$

Taking into account (3.3) one obtains the assertion (4.4). In order to prove (4.1) we insert $f(x) = e^{xt}$ into (4.4). The identity (4.2) follows immediately from (4.1) and (iv). For the proof of (4.3) it is enough to apply the well-known formula for the divided differences with distinct knots

$$[x_0, \cdots, x_n]g = \sum_{j=0}^{n} g(x_j)/w_j.$$

Hence and from (4.1) the desired result follows. This completes the proof. □

COROLLARY 4.1. *Let* $x_j = a + jh$ $(0 \leqq j \leqq n)$. *Then*

$$E(t) = \frac{e^{at}}{n!} \left[\frac{e^{ht} - 1}{h}\right]^n.$$

Proof. In this case we have

$$w_j = (-1)^{n-j} h^n j!(n-j)!.$$

Direct application of (4.3) yields the desired result. □

5. The recurrence formulas for h_r. For the sake of notation we write often $h_r(i, j)$ instead of $h_r(x_i, x_{i+1}, \cdots, x_j)$ $(0 \leqq i \leqq j \leqq n)$. We are ready to state and prove the main result of this section.

THEOREM 5.1. *For any $r \in \mathbb{N}_0$ and $n \in \mathbb{N}$ the complete symmetric functions satisfy the following recurrence relations*:

$$h_r(0, n) = h_r(0, n-1) + x_n h_{r-1}(0, n) \qquad (r \in \mathbb{N}), \tag{5.1}$$

$$h_r(0, n) = \sum_{j=0}^{n} \frac{(-1)^{n-j}}{x_j \cdots x_n} h_{n-j+r+1}(0, j), \tag{5.2}$$

$$h_r(0, n) = \frac{1}{x_n - x_0} \{h_{r+1}(1, n) - h_{r+1}(0, n-1)\}, \tag{5.3}$$

$$h_r(0, n) = \frac{1}{x_n - x_0} \{x_n h_r(1, n) - x_0 h_r(0, n-1)\}, \tag{5.4}$$

$$h_r(x_0 + \gamma, \cdots, x_n + \gamma) = \sum_{l=0}^{r} \binom{n+r}{l} \gamma^l h_{r-l}(x_0, \cdots, x_n) \qquad (\gamma \in \mathbb{R}). \tag{5.5}$$

Proof. The recurrence relation (5.1) has been established by Menon [7]. Below we will present another proof. We need Leibniz' formula for divided differences

$$[x_0, \cdots, x_n](f \cdot g) = \sum_{j=0}^{n} ([x_0, \cdots, x_j]f)([x_j, \cdots, x_n]g), \tag{5.6}$$

where f and g are real-valued functions defined on $[x_0, x_n]$. Making use of (3.3) and (5.6) we obtain

$$\begin{aligned} h_r(0, n) &= [x_0, \cdots, x_n](t^{n+r-1} \cdot t) \\ &= ([x_n]t)([x_0, \cdots, x_n]t^{n+r-1}) + ([x_{n-1}, x_n]t)([x_0, \cdots, x_{n-1}]t^{n+r-1}) \\ &= x_n h_{r-1}(0, n) + h_r(0, n-1). \end{aligned}$$

In a similar fashion we can establish (5.2). For our purposes we need the following identity:

$$[x_0, \cdots, x_j]\frac{1}{t} = (-1)^j \Big/ \prod_{l=0}^{j} x_l \qquad (x_l \neq 0, \text{ all } l), \tag{5.7}$$

which follows immediately from (3.18) and (3.3). From (5.6) and (3.3), we obtain

$$\begin{aligned} h_r(0, n) &= [x_0, \cdots, x_n]\left(\frac{1}{t} \cdot t^{n+r+1}\right) \\ &= \left([x_n]\frac{1}{t}\right)([x_0, \cdots, x_n]t^{n+r+1}) \\ &\quad + \left([x_{n-1}x_n]\frac{1}{t}\right)([x_0, \cdots, x_{n-1}]t^{n+r+1}) + \cdots \\ &\quad + \left([x_0, \cdots, x_n]\frac{1}{t}\right)([x_0]t^{n+r+1}). \end{aligned}$$

Hence and from (3.3) and (5.7) the desired result follows. In order to prove the formula (5.3) we apply the well-known recurrence for the divided differences

$$[x_0, \cdots x_n]t^{n+r} = \frac{1}{x_n - x_0} \{[x_1, \cdots, x_n]t^{n+r} - [x_0, \cdots, x_{n-1}]t^{n+r}\}.$$

Hence, by virtue of (3.3), the assertion (5.3) follows. For the proof of (5.4) we employ the recursion (v) setting $i=0$ and $t_{i+j}=x_j$ $(0\leqq j\leqq n)$. Further, multiplying both sides by t^r and performing integration over $[x_0, x_n]$, we obtain, in view of (3.4),

$$\frac{n-1}{n+r}h_r(0,n)=\frac{1}{x_n-x_0}\{x_nh_r(1,n)-x_0h_r(0,n-1)\}$$
$$-\frac{r+1}{n+r}\frac{1}{x_n-x_0}\{h_{r+1}(1,n)-h_{r+1}(0,n-1)\}.$$

Hence and from (5.3) we get the desired result (5.4) provided $n\geqq 2$. Direct calculations show that (5.4) holds true if $n=1$. We prove now the last statement of our theorem. For $r=0$ the assertion is a trivial one. Assume $r\in\mathbb{N}$. Let $M_{0,n}(\cdot\,|x_0,\cdots,x_n)\equiv M_{0,n}(\cdot)$. It is well known that the following identity

$$M_{0,n}(\cdot\,|x_0+\gamma,\cdots,x_n+\gamma)=M_{0,n}(\cdot-\gamma\,|x_0,\cdots,x_n)$$

holds true for any $\gamma\in\mathbb{R}$. According to (3.4) we obtain

$$h_r(x_0+\gamma,\cdots,x_n+\gamma)=\binom{n+r}{r}\int_{x_0+\gamma}^{x_n+\gamma}M_{0,n}(t|x_0+\gamma,\cdots,x_n+\gamma)t^r\,dt$$
$$=\binom{n+r}{r}\int_{x_0+\gamma}^{x_n+\gamma}M_{0,n}(t-\gamma|x_0,\cdots,x_n)t^r\,dt$$
$$=\binom{n+r}{r}\int_{x_0}^{x_n}M_{0,n}(z|x_0,\cdots,x_n)(z+\gamma)^r\,dz,$$

where $z=t-\gamma$. Hence and from (3.4) the assertion follows. The proof is completed. □

6. Inequalities involving h_r. In this section we assume $x_j\geqq 0$, all j. Thus $0\leqq x_0\leqq\cdots\leqq x_n$ with $x_0<x_n$. In [7] it is proved that

(6.1) $$h_{p-m}h_{q+m}\geqq h_{p-m-1}h_{q+m+1} \qquad (p\leqq q, 0\leqq m<p),$$

(6.2) $$h_{r-1}h_{r+1}\leqq h_r^2 \qquad (r\in\mathbb{N}),$$

(6.3) $$h_r^{1/r}\geqq h_s^{1/s} \qquad (1\leqq r\leqq s).$$

These inequalities are strict unless all but one of the variables are zero. The inequality (6.2) tells us that the sequence $\{h_r\}_0^\infty$ is logarithmically concave. The companion inequality to (6.2) is provided by the following.

THEOREM 6.1. *Let $p, q, r>0$ and let $1/p+1/q=1/r$. Then the following inequality*

(6.4)
$$\left[\binom{n+r(l+m)}{n}^{-1}h_{r(l+m)}\right]^{1/r}\leqq\left[\binom{n+pl}{n}^{-1}h_{pl}\right]^{1/p}\cdot\left[\binom{n+qm}{n}^{-1}h_{qm}\right]^{1/q}$$

holds true provided that $r(l+m)$, pl, $qm\in\mathbb{N}_0$. Hence, in particular

(6.5)
$$\left[\binom{n+l+m}{n}^{-1}h_{l+m}\right]^2\leqq\left[\binom{n+2l}{n}^{-1}h_{2l}\right]\cdot\left[\binom{n+2m}{n}^{-1}h_{2m}\right] \qquad (l+m, 2l, 2m\in\mathbb{N}_0),$$

and

(6.6) $$h_r^2\leqq\frac{(n+r)(r+1)}{(n+r+1)r}h_{r-1}h_{r+1} \qquad (r\in\mathbb{N}).$$

Proof. Let $L_p \equiv L_p(S, \Sigma, \mu)$, $-\infty < p < \infty$, be the space of all pth power nonnegative integrable functions over a given finite measure space (S, Σ, μ). If $f \in L_p$ and if

$$\|f\|_p = \left[\int_S f^p \, d\mu\right]^{1/p},$$

then for any $f_1 \in L_p$ and any $f_2 \in L_q$ with $1/p + 1/q = 1/r$ $(p, q, r > 0)$

$$\|f_1 \cdot f_2\|_r \leqq \|f_1\|_p \|f_2\|_q$$

(see [16]). In order to prove the inequality (6.4) we set $d\mu = dt$, $f_1(t) = M_{0,n}(t)^{1/p} t^l$, $f_2(t) = M_{0,n}(t)^{1/q} \cdot t^m$. Taking into account the above result and (3.4), we arrive at (6.4). The inequality (6.5) follows immediately from (6.4) by letting $p = q = 2$. Setting $l = (r+1)/2$, $m = (r-1)/2$ into (6.5) we obtain (6.6). This completes the proof. □

Let

$$A_r = \left[\frac{1}{n+1} \sum_{i=0}^{n} x_i^r\right]^{1/r} \qquad (r \in \mathbb{R};\ r \neq 0)$$

denote the rth mean in variables $x_0, \cdots, x_n$. It is well known that $A_1 \leqq A_r (r \geqq 1)$. We are in a position to prove the following.

THEOREM 6.2. *For any* $r \in \mathbb{N}$

$$A_1 \leqq \left[\binom{n+r}{r}^{-1} h_r\right]^{1/r} \leqq A_r. \tag{6.7}$$

Proof. In order to prove the inequality (6.7) we will apply the following result:

$$f(A_1) \leqq \int_{x_0}^{x_n} M_{0,n}(t) f(t)\, dt \leqq \frac{1}{n+1} \sum_{j=0}^{n} f(x_j),$$

where f is a convex function on (x_0, x_n). The inequalities are strict unless f is a polynomial of degree one or less (see [10]). Setting above $f(t) = t^r$ $(r \in \mathbb{N})$ and next making use of (3.4) we arrive at (6.7). The proof is completed. □

Our next result reads as follows.

THEOREM 6.3. *Let* $m, r \in \mathbb{N}_0$ *and let*

$$\alpha \in \begin{cases} [x_n, \infty) & \text{if } m \text{ is odd,} \\ (-\infty, \infty) & \text{if } m \text{ is even.} \end{cases}$$

Then

$$\sum_{j=0}^{m} (-1)^{m-j} \binom{m}{j} \binom{n+m+r-j}{n}^{-1} \alpha^j h_{m+r-j} \geqq 0. \tag{6.8}$$

Proof. There is nothing to prove when $m = 0$. Assume $m > 0$. The inequality (6.8) follows immediately from the obvious inequality

$$\int_{x_0}^{x_n} t^r (\alpha - t)^m M_{0,n}(t)\, dt \geqq 0$$

and from (3.4). This completes the proof. □

Imposing some restrictions on the distribution of the x's we can prove more inequalities involving h_r.

THEOREM 6.4. *Let* $1/(n+1)\sum_{j=0}^{n} x_j \geqq 1$. *Then for any* r *and* s *with* $0 \leqq r \leqq s$ *the following inequalities*

$$h_r \leqq \frac{(r+1)(r+2)\cdot\cdots\cdot s}{(n+r+1)(n+r+2)\cdot\cdots\cdot(n+s)} h_s, \tag{6.9}$$

and

$$h_r \leqq h_s \tag{6.10}$$

hold true. Moreover, if $0 \leqq x_0 \leqq \cdots \leqq x_n \leqq 1$ *with* $x_0 < x_n$, *then*

$$h_r \geqq \frac{(r+1)(r+2)\cdot\cdots\cdot s}{(n+r+1)(n+r+2)\cdot\cdots\cdot(n+s)} h_s. \tag{6.11}$$

Proof. Let $H_r := \binom{n+r}{r}^{-1} h_r$ $(r \in \mathbb{N}_0)$. Hence in particular $H_0 = 1$, $H_1 = 1/(n+1)\sum_{j=0}^{n} x_j$. It has been shown in [9] that the sequence $\{h_r\}_0^\infty$ is logarithmically convex, i.e.,

$$H_r^2 \leqq H_{r-1} H_{r+1} \qquad (r \in \mathbb{N}). \tag{6.12}$$

We already know that the sequence $\{h_r\}_0^\infty$ is logarithmically concave. In order to prove the inequality (6.9) let us observe that if for some $r \in \mathbb{N}$, $H_{r-1} \leqq H_r$, then also $H_r \leqq H_{r+1}$. This follows from (6.12). The inequality $H_r \leqq H_{r+1} (r \in \mathbb{N}_0)$ implies

$$h_r \leqq \frac{r+1}{n+r+1} h_{r+1}.$$

Hence (6.9) follows. Inequality (6.10) is an obvious consequence of the previous inequality. For the proof of (6.11) we apply Theorem 6.3 with $m = \alpha = 1$. The proof is completed. □

We close this section with the following theorem.

THEOREM 6.5. *Let*

$$0 \leqq x_0 \leqq \cdots \leqq x_{i-1} < x_i = \cdots = x_n$$

for some i $(1 \leqq i \leqq n)$. *Then*

$$h_{n+r-i}(x_0, \cdots, x_i) \leqq \frac{x_n^{n+r}}{(x_n - x_0)\cdot\cdots\cdot(x_n - x_{i-1})}. \tag{6.13}$$

In particular if $0 \leqq x_0 \leqq \cdots \leqq x_{n-1} < x_n$, *then*

$$h_r(x_0, \cdots, x_n) \leqq \frac{x_n^{n+r}}{(x_n - x_0)\cdot\cdots\cdot(x_n - x_{n-1})}. \tag{6.14}$$

Proof. Let f be a real-valued and sufficiently smooth function defined on $[x_0, x_n]$. Then Newton's theorem provides

$$f(t) = \sum_{j=0}^{n+r} ([x_0, \cdots, x_j]f) \prod_{l=0}^{j-1} (t - x_l) + R(f),$$

where $R(f) = 0$ if and only if f is a polynomial of degree $n+r$ or less. Setting above $f(t) = t^{n+r}$ and next using (3.3), we obtain

$$t^{n+r} = \sum_{j=0}^{n+r} h_{n+r-j}(0, j) \prod_{l=0}^{j-1} (t - x_l).$$

Putting $t = x_n$, we get

$$x_n^{n+r} = \sum_{j=0}^{n+r} h_{n+r-j}(0, j) \prod_{l=0}^{j-1} (x_n - x_l).$$

If for some i $(1 \leqq i \leqq n)$, $x_i = \cdots = x_n$, then the above identity reduces to

$$x_n^{n+r} = h_{n+r-i}(0, i) \prod_{l=0}^{i-1} (x_n - x_l) + \sum_{j=0}^{i-1} h_{n+r-j}(0, j) \prod_{l=0}^{j-1} (x_n - x_l).$$

Since the last sum is nonnegative, the assertion (6.13) follows. Setting $i = n$ in (6.13), we obtain (6.14). This completes the proof. □

7. Applications. In this section we give some two-fold applications of our previous results. In §§ 7.1 and 7.2 we deal with the q-binomial coefficients and the r-Stirling numbers of the second kind, respectively.

7.1. Assume $x_i = q^i$ $(i = 0, 1, \cdots, n;\ q > 0,\ q \neq 1)$. We already know that $h_r = \begin{bmatrix} n+r \\ r \end{bmatrix}$ $(r \in \mathbb{N}_0;\ n \in \mathbb{N})$. Hence and from (5.1) we rediscover the well-known recurrence formula for the q-binomial coefficients

$$\begin{bmatrix} m \\ r \end{bmatrix} = \begin{bmatrix} m-1 \\ r \end{bmatrix} + q^{m-r} \begin{bmatrix} m-1 \\ r-1 \end{bmatrix} \qquad (1 \leqq r \leqq m)$$

(see, e.g., [1]). From (5.2), we obtain in a similar manner

$$\begin{bmatrix} m \\ r \end{bmatrix} = \sum_{j=0}^{m-r} \frac{(-1)^{m-r-j}}{q^j \cdots q^{m-r}} \begin{bmatrix} m+1 \\ j \end{bmatrix}.$$

From (6.4), we obtain the following inequality involving the q-binomial coefficients:

$$\left\{ \binom{n+r(l+m)}{n}^{-1} \begin{bmatrix} n+r(l+m) \\ n \end{bmatrix} \right\}^{1/r}$$
$$\leqq \left\{ \binom{n+pl}{n}^{-1} \begin{bmatrix} n+pl \\ n \end{bmatrix} \right\}^{1/p} \left\{ \binom{n+qm}{n}^{-1} \begin{bmatrix} n+qm \\ n \end{bmatrix} \right\}^{1/q}$$
$$(p, q, r > 0;\ 1/p + 1/q = 1/r;\ (l+m)r, pl, qm \in \mathbb{N}_0).$$

Combining (6.2) and (6.6) we get

$$\begin{bmatrix} m-1 \\ r-1 \end{bmatrix} \begin{bmatrix} m+1 \\ r+1 \end{bmatrix} \leqq \begin{bmatrix} m \\ r \end{bmatrix}^2 \leqq \frac{m(r+1)}{(m+1)r} \begin{bmatrix} m-1 \\ r-1 \end{bmatrix} \begin{bmatrix} m+1 \\ r+1 \end{bmatrix} \qquad (r, m \in \mathbb{N};\ r \leqq m).$$

The inequality (6.3) gives us

$$\begin{bmatrix} n+r \\ r \end{bmatrix}^{1/r} \geqq \begin{bmatrix} n+s \\ s \end{bmatrix}^{1/s} \qquad (1 \leqq r \leqq s;\ n \in \mathbb{N}_0).$$

The inequality (6.7) implies the following inequality:

$$\frac{q^{n+1}-1}{(n+1)(q-1)} \leqq \left\{ \binom{n+r}{r}^{-1} \begin{bmatrix} n+r \\ r \end{bmatrix} \right\}^{1/r} \leqq \left\{ \frac{q^{r(n+1)}-1}{(n+1)(q-1)} \right\}^{1/r}$$
$$(r \in \mathbb{N};\ n \in \mathbb{N}_0).$$

From (6.9), we obtain

$$\begin{bmatrix} n+r \\ r \end{bmatrix} \leqq \frac{(r+1)(r+2) \cdot \cdots \cdot s}{(n+r+1)(n+r+2) \cdot \cdots \cdot (n+s)} \begin{bmatrix} n+s \\ s \end{bmatrix} \qquad (0 \leqq r \leqq s)$$

provided that $q > 1$. This inequality is reversed if $0 < q < 1$. Finally, (6.14) implies the following two inequalities:

$$\begin{bmatrix} n+r \\ r \end{bmatrix} \leqq \frac{1}{(1-q^n) \cdot \cdots \cdot (1-q)}$$

when $0<q<1$ and

$$\begin{bmatrix} n+r \\ r \end{bmatrix} \leqq \frac{q^{n(n+r)}}{(q^n-1)\cdot\cdots\cdot(q^n-q^{n-1})}$$

when $q>1$.

7.2. Assume now $x_i = r+i$ with $r\in\mathbb{N}_0$ and $i=0, 1, \cdots, n$. First we will show that the identity (1.2) holds true. It is proved that

$$S_r(k+r, n+r) = [r, r+1, \cdots, r+n]t^k$$

($r\in\mathbb{N}_0$, $k\geqq n\geqq 0$) (cf. [3]). According to (3.3) we can rewrite the last identity as

$$S_r(k+r, n+r) = h_{k-n}(r, r+1, \cdots, r+n).$$

Hence (1.2) follows. Applying (1.2) to (5.1), we obtain

$$S_r(k, n) = S_r(k, n-1) + nS_r(k-1, n) \qquad (0\leqq r\leqq k;\ k\geqq n-1)$$

(see [3]). Equation (5.2) leads to the following recurrence relation for the r-Stirling numbers of the second kind

$$S_r(k+r, n+r) = \sum_{j=0}^{n} \frac{(-1)^{n-j}}{(r+j)\cdot\cdots\cdot(r+n)} S_r(k+r+1, j+r) \qquad (0\leqq r\leqq k;\ n\in\mathbb{N}_0).$$

Setting $\gamma = r$, $x_i = i$ $(0\leqq i\leqq n)$ into (5.5), we obtain

$$S_r(n+k, n+r) = \sum_{m=n}^{k} \binom{k}{m} S(m, n) r^{k-m} \qquad (0\leqq n\leqq k;\ r\in\mathbb{N}_0),$$

where, as usual, $S(\cdot, \cdot)$ denotes the Stirling number of the second kind. The polynomial (in r) that appears on the right-hand side is commonly referred to as the Stirling polynomial of the second kind (see [4]). We will close this section giving three inequalities involving the numbers $S_r(\cdot, \cdot)$. Combining (6.2) and (6.6) gives us

$$S_r(l-1, k)S_r(l+1, k) \leqq S_r(l, k)^2$$

$$\leqq \frac{(l-r)(l-k+1)}{(l-r+1)(l-k)} S_r(l-1, k)\cdot S_r(l+1, k) \qquad (0\leqq r\leqq k<l).$$

The left inequality tells us that the sequence $\{S_r(\cdot, k)\}$ is logarithmically concave. Our second inequality reads as follows:

$$\frac{(r+n)^k - n(r+n-1)^k}{n!} \leqq S_r(k+r, n+r) \leqq \frac{(r+n)^k}{n!} \qquad (r\in\mathbb{N}_0;\ 0\leqq n\leqq k).$$

There is nothing to prove when $n=0$. Assume $n>0$. In this case the right inequality follows from (6.10) and (1.2). For the proof of the left inequality it is enough to apply the following one:

$$[r, r+1, \cdots, r+n]t^k \geqq (r+n)^k/n! - (r+n-1)^k/(n-1)!$$

(cf. [17]). Next, taking into account that

$$[r, r+1, \cdots, r+n]t^k = S_r(k+r, n+r) \qquad (r\in\mathbb{N}_0;\ k\geqq n\geqq 0)$$

we obtain the desired result. Our last inequality reduces to that given by Wegner [17] (see also [15]) when $r=0$.

Before the presentation of our last result we introduce more notation. Let $0 \leqq \mu \leqq \nu$ (μ, ν-integers) and let

$$x(t) = \sum_{m=\mu}^{\nu} a_m t^m \qquad (a_m, t \in \mathbb{R}).$$

Further, let

$$a = \min\{x(t): r \leqq t \leqq r+n\},$$
$$b = \max\{x(t): r \leqq t \leqq r+n\},$$

with $r \in \mathbb{N}_0$, $n = \mathbb{N}$. If f is a convex function on (a, b), then the following inequality

$$f\left[\sum_{m=\mu}^{\nu} a_m \binom{m+n}{n}^{-1} S_r(m+n+r, n+r)\right] \leqq \left[J_n \int_r^{r+n} f(x(t))^2\, dt\right]^{1/2}$$

holds true, where

$$J_n = 2n \sum_{j=1}^{n} (-1)^{n-j} \frac{j^{2n-1}}{(n-j)!(n+j)!} \qquad (n \in \mathbb{N}).$$

Our inequality is a special case of the following one:

$$g\left[\sum_{m=\mu}^{\nu} a_m \binom{m+n}{n}^{-1} h_m(x_0, \cdots, x_n)\right] \leqq \left[J_n \int_{x_0}^{x_n} g(x(t))^2\, dt\right]^{1/2}$$

with g convex on (a, b), where now

$$a = \min\{x(t): x_0 \leqq t \leqq x_n\},$$
$$b = \max\{x(t): x_0 \leqq t \leqq x_n\}$$

(see [9] for the details). Setting $x_j = j + r$, all j, we obtain the desired result.

Let us notice that $J_1 = 1$, $J_2 = 2/3$, $J_3 = 11/20$, $J_4 = 151/315$, $J_5 = 15619/36288$.

Acknowledgments. The author thanks Professor Stefan Paszkowski for his proofs of Corollaries 3.1 and 3.2. Thanks are due to Professor Philip Feinsilver for drawing my attention to the Feynman identity and for useful conversations. I would also like to thank Mrs. Linda Macak for her excellent typing.

REFERENCES

[1] G. A. ANDREWS, *The theory of partitions*, in Encyclopedia of Mathematics, Vol. 2, Addison-Wesley, London, 1976.
[2] K. E. ATKINSON, *An Introduction to Numerical Analysis*, John Wiley, New York, 1978.
[3] A. BRODER, *The r-Stirling numbers*, Discrete Math., 49 (1984), pp. 241-259.
[4] L. CARLITZ, *Weighted Stirling numbers of the first and second kind*—I, Fibonacci Quart., 18 (1980), pp. 147-162.
[5] H. B. CURRY AND I. J. SCHOENBERG, *On Pólya frequency functions* IV. *The fundamental spline functions and their limits*, J. Anal. Math., 17 (1966), pp. 71-107.
[6] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Clarendon Press, Oxford, 1979.
[7] K. V. MENON, *Inequalities for symmetric functions*, Duke Math. J., 35 (1968), pp. 37-45.
[8] L. M. MILNE-THOMSON, *The Calculus of Finite Differences*, Macmillan, London, 1951.
[9] E. NEUMAN, *Inequalities involving generalized symmetric means*, J. Math. Anal. Appl., 120 (1986), pp. 315-320.
[10] ———, *On interpolating means*, J. Math. Anal. Appl., to appear.
[11] S. PASZKOWSKI, *Collection of Exercises in Numerical Analysis. Part* I, Polish Scientific Publishers, Lodz, 1969. (In Polish.)

[12] S. PASZKOWSKI, *Numerical Applications of the Chebyshev Polynomials and Series*, Polish Scientific Publishers, Warsaw, 1975. (In Polish.)
[13] ———, private communication.
[14] L. L. SCHUMAKER, *Spline Functions: Basic Theory*, John Wiley, New York, 1981.
[15] M. SOBEL, V. R. R. UPPULURI, AND K. FRANKOWSKI, *Selected Tables in Mathematical Statistics, Vol.* 4, American Mathematical Society, Providence, RI, 1977.
[16] CH. L. WANG, *Variants of the Hölder inequality and its inverses*, Canad. Math. Bull., 20 (1977), pp. 377–384.
[17] H. WEGNER, *Über die Stirlingschen Zahlen der zweiter Art*, J. Reine Angew. Math., 266 (1974), pp. 88–99.

SIAM J. MATH. ANAL.
Vol. 19, No. 3, May 1988

017

OSCILLATION PROPERTIES FOR SOME POLYNOMIAL ANALOGUES OF THE PROLATE SPHEROIDAL WAVE FUNCTIONS*

MARCI A. PERLSTADT†

Abstract. Slepian, Landau, and Pollak found that a certain finite integral operator commutes with a much simpler second-order differential operator. The eigenfunctions that these operators share are prolate spheroidal wave functions and the study of these eigenfunctions has led to applications in several areas. Grünbaum displayed analogues of this commutativity for certain integral operators involving orthogonal polynomials. We discuss some implications of this commutativity for these eigenfunctions.

Key words. orthogonal polynomials, prolate spheroidal wave functions, oscillation

AMS(MOS) subject classifications. primary 33A65; secondary 42A16

1. Introduction. Let f be a square integrable function with Fourier transform $\hat{f} = Ff$. For $\mathscr{A}$ and $\mathscr{B}$ subsets of $\mathbb{R}$, we let A be the operator that restricts f (timelimits f) to $\mathscr{A}$ and let B be the operator that restricts $\hat{f}$ (bandlimits f) to $\mathscr{B}$, i.e.,

$$Af = f \cdot \chi_{\mathscr{A}} \quad \text{and} \quad B\hat{f} = \hat{f} \cdot \chi_{\mathscr{B}}.$$

In a series of papers [1], [2], [3], Slepian, Landau, and Pollak studied the integral operator that timelimits then bandlimits and then again timelimits a function, $AF^{-1}BFA$ (here F^{-1} denotes the inverse Fourier transform). The eigenvalues and eigenfunctions of this operator proved critical to an understanding of the "space" of "nearly" time and bandlimited functions. The realization that the operator $AF^{-1}BFA$ commuted with a relatively simple Sturm–Liouville type second-order differential equation with simple spectrum [1] was crucial in determining many properties of the eigenfunctions, including the fact that the eigenfunctions were prolate spheroidal wave functions.

In recent years the problem has been looked at for more general Fourier situations [4], [5]. In particular, we will be interested in the case of expansions in terms of orthogonal polynomials $\{p_i(x)\}$ where the orthogonality is with respect to a nonnegative continuous weight $w(x)$ on $\mathscr{C}$. Assuming polynomials are complete for the space of functions square integrable with respect to the inner product

$$\langle f, g\rangle_{w(x),\mathscr{C}} = \int_{\mathscr{C}} f(x)g(x)w(x)\,dx,$$

we have for such functions that

$$f(x) = \sum_{i=0}^{\infty} \hat{f}(i)p_i(x) \quad \text{where } F(f) = \hat{f}(i) = \langle f(x), p_i(x)\rangle_{w(x),\mathscr{C}}.$$

Here and throughout the remainder of this paper we assume the $p_i(x)$'s are orthonormal.

If now A is the operator that restricts f to $\mathscr{A} \subset \mathscr{C}$, i.e.,

$$Af = f \cdot \chi_{\mathscr{A}}$$

and B is the operator that restricts $\hat{f}(i)$ via

$$B\hat{f}(i) = \begin{cases} \hat{f}(i), & i = 0, 1, \cdots, L, \\ 0, & i > L, \end{cases}$$

* Received by the editors October 29, 1986; accepted for publication (in revised form) May 18, 1987.
† Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

we can again study $AF^{-1}BFA$. In this case (see [6])

$$(1.1)\quad AF^{-1}BFAf(x)=\int_{A} K_L(x,y)f(y)w(y)\,dy, \qquad K_L(x,y)=\sum_{i=0}^{L} p_i(x)p_i(y).$$

We can also study the band-time-bandlimiting operator $BFAF^{-1}B$ which can be represented as an $(L+1)\times(L+1)$ matrix G [6] with

$$(1.2)\quad (G)_{ij}=\int_{\mathscr{A}} p_i(x)p_j(x)w(x)\,dx, \qquad i,j=0,1,\cdots,L.$$

In [5] it is shown that for the classical orthogonal polynomials, $AF^{-1}BFA$ has a commuting second-order differential operator and G has a commuting tridiagonal matrix provided $\mathscr{A}$ is chosen properly. We will use this commutativity to establish a number of properties of the eigenfunctions for these operators. Most of these properties are analogous to those found in [1] for the case of the standard Fourier transform and these eigenfunctions can be viewed as polynomial analogues of the prolate spheroidal wave functions.

2. Background. We begin by noting the following lemma whose proof can be found in [6].

LEMMA 2.1. *Let $\{p_i(x)\}$ be a complete orthonormal family of polynomials ($i=0,1,2,\cdots$, degree of $p_i(x)=i$) with respect to the continuous nonnegative weight function $w(x)$ on $\mathscr{C}$. Let $\mathscr{A}\subset\mathscr{C}$ be such that*

$$0<\int_{\mathscr{A}} w(x)\,dx<\int_{\mathscr{C}} w(x)\,dx.$$

Then

(i) *$AF^{-1}BFA$ and $BFAF^{-1}B$ have $L+1$ positive eigenvalues,*

$$1>\lambda_0\geqq\lambda_1\geqq\cdots\geqq\lambda_L>0.$$

(ii) *If $\phi_0(x),\phi_1(x),\cdots,\phi_L(x)$ are eigenfunctions of $AF^{-1}BFA$ corresponding to eigenvalues $\lambda_0,\lambda_1,\cdots,\lambda_L$, respectively, then $BF\phi_i$ is an eigenvector of $BFAF^{-1}B$ corresponding to λ_i for $i=0,1,\cdots,L$.*

(iii) *If $\vec{c}^{(0)},\vec{c}^{(1)},\cdots,\vec{c}^{(L)}$ are eigenvectors of $BFAF^{-1}B$ corresponding to $\lambda_0,\lambda_1,\cdots,\lambda_L$, respectively, then $AF^{-1}\vec{c}^{(i)}$ is an eigenfunction of $AF^{-1}BFA$ corresponding to λ_i.*

(iv) *The $\phi_i(x)$'s are polynomials of degree less than or equal to L and can be normalized so that*

(a) $\lambda_i\phi_i(x)=\int_{\mathscr{A}} K_L(x,y)\phi_i(y)w(y)\,dy$,

(b) $\langle\phi_i(x),\phi_j(x)\rangle_{w(x),\mathscr{C}}=\delta_{ij}$, *and*

(c) $\langle\phi_i(x),\phi_j(x)\rangle_{w(x),\mathscr{A}}=\lambda_i\delta_{ij}$.

(v) *If we let $\sim\mathscr{A}=\mathscr{C}-\mathscr{A}$ and $\sim Af=f\cdot\chi_{\sim\mathscr{A}}$ we have*

(a) $\sim AF^{-1}BF\sim A\phi_i(x)=(1-\lambda_i)\phi_i(x)$, *and*

(b) $\langle\sim A\phi_i,\sim A\phi_i\rangle_{w(x),\mathscr{C}}=\langle\phi_i,\phi_i\rangle_{w(x),\sim\mathscr{A}}=1-\lambda_i$.

We now turn our attention to the problem Grünbaum investigated in [5], namely that of determining for which orthogonal polynomial families the operator $AF^{-1}BFA$ commutes with a second-order differential operator and the matrix G of (1.2) commutes with a tridiagonal matrix. He found the following.

THEOREM 2.2. *For the classical orthogonal polynomials (Jacobi, Hermite, Laguerre, and Bessel), orthogonal on (δ,ε) with respect to $w(x)$, if we choose $A=(\sigma,\varepsilon)$ where*

$\delta < \sigma < \varepsilon$, then there exists a second-order differential operator D commuting with $AF^{-1}BFA$ and a tridiagonal matrix T commuting with G.

The differential operators D are explicitly constructed in [5]. If we apply D to $p_n(x)$ and use a number of properties peculiar to the classical orthogonal polynomials, we find that there are constants T_{ij} such that

$$Dp_n(x) = T_{nn-1}p_{n-1}(x) + T_{nn}p_n(x) + T_{nn+1}p_{n+1}(x).$$

Thus taking

$$\phi_i(x) = \sum_{n=0}^{L} c_n^{(i)} p_n(x) \tag{2.3}$$

and applying D to (2.3), we obtain a three-term recurrence for the $c_n^{(i)}$:

$$T_{nn-1}c_{n-1}^{(i)} + T_{nn}c_n^{(i)} + T_{nn+1}c_{n+1}^{(i)} = \mu_i c_n^{(i)}$$

where $D\phi_i(x) = \mu_i\phi_i(x)$. Thus we may take the T_{ij}'s as the entries of the matrix T that commutes with G. An example of this sort is carried out in [4].

Our efforts will be limited to the Jacobi, Hermite, and Laguerre polynomials. Since we will make use of D and T, we will give them explicitly below.

The orthonormalized polynomials $\{p_i(x)\}$ all satisfy [9] recurrences of the form

$$b(n)xp_n(x)\sqrt{h_n} = a(n)\sqrt{h_{n+1}}\,p_{n+1}(x) + c(n)\sqrt{h_n}\,p_n(x) + d(n)\sqrt{h_{n-1}}\,p_{n-1}(x),$$

second-order differential equations of the form

$$\frac{1}{w(x)}\frac{d}{dx}\left[W(x)\frac{d}{dx}p_n(x)\right] = \mu_n p_n(x),$$

and first-order equations of the form

$$\frac{W(x)}{w(x)}\frac{d}{dx}p_n(x) = \alpha_n p_n(x) + x\beta_n p_n(x) + \sqrt{\frac{h_{n-1}}{h_n}}\,\gamma_{n-1}p_{n-1}(x).$$

Here we have (following the notation in [9]) that for the

(a) Jacobi polynomials on $(-1, 1)$:

$w(x) = (1-x)^\alpha(1+x)^\beta \quad (\alpha, \beta > -1)$, $\qquad W(x) = (1-x^2)w(x)$,

$\mu_n = -n(n+\alpha+\beta+1)$, $\qquad h_n = \dfrac{2^{\alpha+\beta+1}}{\alpha+\beta+2n+1} \cdot \dfrac{\Gamma(\alpha+n+1)\Gamma(\beta+n+1)}{n!\Gamma(\alpha+\beta+n+1)}$,

$a(n) = 2(n+1)(\alpha+\beta+n+1)(\alpha+\beta+2n)$, $\qquad b(n) = (\alpha+\beta+2n)(\alpha+\beta+2n+1) \cdot (\alpha+\beta+2n+2)$,

$c(n) = (\alpha+\beta+2n+1)(\beta^2-\alpha^2)$, $\qquad d(n) = 2(\alpha+n)(\beta+n) \cdot (\alpha+\beta+2n+2)$,

$\alpha(n) = \dfrac{n(\alpha-\beta)}{\alpha+\beta+2n}$, $\qquad \beta(n) = -n$,

$\gamma(n-1) = \dfrac{2(n+\alpha)(n+\beta)}{\alpha+\beta+2n}$;

(b) Laguerre polynomials on $(0, \infty)$:

$$w(x) = e^{-x}x^{\alpha} \quad (\alpha > -1), \qquad W(x) = e^{-x}x^{\alpha+1},$$

$$\mu_n = -n, \qquad h_n = \Gamma(1+\alpha)\binom{n+\alpha}{n},$$

$$a(n) = -(n+1), \qquad b(n) = 1,$$

$$c(n) = (2n+\alpha+1), \qquad d(n) = -(n+\alpha),$$

$$\alpha(n) = n, \qquad \beta(n) = 0,$$

$$\gamma(n-1) = -(n+\alpha);$$

(c) Hermite polynomials on $(-\infty, \infty)$:

$$w(x) = e^{-x^2}, \qquad W(x) = e^{-x^2},$$

$$\mu_n = -2n, \qquad h_n = \pi^{1/2}2^n n!,$$

$$a(n) = 1, \qquad b(n) = 2,$$

$$c(n) = 0, \qquad d(n) = 2n,$$

$$\alpha(n) = 0, \qquad \beta(n) = 0,$$

$$\gamma(n-1) = 2n.$$

The commuting operators D, given in [5], have the form

$$D = \frac{1}{w(x)}\frac{d}{dx}\left[(x-\sigma)W(x)\frac{d}{dx}\right] + Ax$$

where for the Jacobi polynomials, $A = L(L+\alpha+\beta+2)$; Laguerre polynomials, $A = L$; Hermite polynomials, $A = 2L$.

The matrices T are given by

$$T_{i+1,i} = T_{i,i+1} = (A + \mu_i + \beta_i)\frac{a(i)}{b(i)}\sqrt{\frac{h_{i+1}}{h_i}}, \qquad i = 0, 1, \cdots, L-1,$$

$$T_{i,i} = (A + \mu_i + \beta_i)\frac{c(i)}{b(i)} + \alpha_i - \sigma\mu_i, \qquad i = 0, 1, \cdots, L.$$

Note that since all of the super- and subdiagonal elements of the matrices T are nonzero, we are guaranteed that T has simple spectrum [10, p. 300]. Thus the eigenvectors of T are also eigenvectors of G.

3. Statement of the main result. Our goal here is to show that in the special cases where commuting operators D and T exist, we know a great deal more about the eigenfunctions ϕ_i and eigenvectors $\vec{c}^{(i)}$ than what was stated in Lemma 2.1. We have, in fact, the following.

THEOREM 3.1. *For the Jacobi, Hermite, and Laguerre polynomials with $A = (\sigma, \varepsilon)$ as described above we have*

(a) *$AF^{-1}BFA$ and $BFAF^{-1}B$ have simple spectrum (away from 0), i.e., $1 > \lambda_0 > \cdots > \lambda_L > 0$.*

(b) *If the eigenvalues of the commuting matrices T are ordered so that $\mu_0 > \mu_1 > \cdots > \mu_L$ with corresponding eigenfunctions $\phi_{i_0}\phi_{i_1}, \cdots, \phi_{i_L}$ then, in fact, $i_0 = 0, i_1 = 1, \cdots, i_L = L$.*

(c) *$\phi_0, \phi_1, \cdots, \phi_L$ each have L simple zeros on $\mathscr{C} = (\delta, \varepsilon)$ and ϕ_i has exactly i zeros on $A = (\sigma, \varepsilon)$. Furthermore the zeros of ϕ_i separate those of ϕ_{i+1}.*

(d) *For a vector $\vec{x} = (x_0, \cdots, x_L)$, let $S^-(\vec{x})$ be the number of sign changes in the sequence $x_0, x_1, \cdots, x_L$ with zero terms discarded. Let $S^+(\vec{x})$ be the maximum number of sign changes in the sequence $x_0, x_1, \cdots, x_L$ where zero terms are arbitrarily assigned values of $+1$ or -1. Then $S^-(\vec{c}^{(i)}) = S^+(\vec{c}^{(i)}) = i$, for $i = 0, 1, \cdots, L$, for the case of Jacobi and Hermite polynomials. For the Laguerre polynomials $S^-(\vec{c}^{(i)}) = S^+(\vec{c}^{(i)}) = L - i$.*

Before beginning the proof of Theorem 3.1 we note that in many applications we are primarily interested in the $\phi_i(\vec{c}^{(i)})$ as eigenfunctions (eigenvectors) of $AF^{-1}BFA$ ($BFAF^{-1}B$) as opposed to as eigenfunctions (eigenvectors) of $D(T)$. Theorem 3.1(b) allows us to extend properties of the ϕ_i and $\vec{c}^{(i)}$ that are readily derived from D and T to the operators $AF^{-1}BFA$ and $BFAF^{-1}B$.

Proof. We begin by showing that if (a) and (b) are established then (c) and (d) follow readily.

If we write

$$\text{(3.1(e))} \qquad \frac{d}{dx}\left((x-\sigma)W(x)\frac{d}{dx}\phi_n(x)\right) + Aw(x)x\phi_n(x) = \mu_n w(x)\phi_n(x),$$

then we can apply the argument in [7, pp. 719–721] to show that if $\mu_i > \mu_j$, then ϕ_j has at least one more zero than ϕ_i on (σ, ε). For completeness we sketch the argument here.

Suppose that $\mu_i > \mu_j$. Multiplying equation (3.1(e)) for ϕ_i by ϕ_j and equation (3.1(e)) for ϕ_j by ϕ_i and subtracting yields

$$\text{(3.1(f))} \qquad \begin{aligned} &\frac{d}{dx}\left((x-\sigma)W(x)\left(\phi_j(x)\frac{d}{dx}\phi_i(x) - \phi_i(x)\frac{d}{dx}\phi_j(x)\right)\right) \\ &\qquad = (\mu_i - \mu_j)w(x)\phi_i(x)\phi_j(x). \end{aligned}$$

If we integrate this equation from σ to $\tilde{x}$ ($\tilde{x} < \varepsilon$), we get

$$\text{(3.1(g))} \qquad (\tilde{x}-\sigma)W(\tilde{x})(\phi_j(\tilde{x})\phi_i'(\tilde{x}) - \phi_i(\tilde{x})\phi_j'(\tilde{x})) = (\mu_i - \mu_j)\int_\sigma^{\tilde{x}} \phi_i(x)\phi_j(x)w(x)\,dx.$$

Suppose $\tilde{x}$ is chosen so that $\phi_i(\tilde{x}) = 0$ and $\phi_i(x) \neq 0$, $x \in (\sigma, \tilde{x})$. Without loss of generality we may assume $\phi_i(x) > 0$ on $(\sigma, \tilde{x})$ and thus that $\phi_i'(\tilde{x}) < 0$. Suppose now that $\phi_j(x)$ has no zeros in $(\sigma, \tilde{x})$ and without loss of generality we take $\phi_j(x) > 0$ on $(\sigma, \tilde{x})$. Then (3.1(g)) becomes

$$\text{(3.1(h))} \qquad (\tilde{x}-\sigma)W(\tilde{x})\phi_j(\tilde{x})\phi_i'(\tilde{x}) = (\mu_i - \mu_j)\int_\sigma^{\tilde{x}} \phi_i(x)\phi_j(x)w(x)\,dx.$$

But in all cases being studied $(\tilde{x}-\sigma)W(\tilde{x}) > 0$ and $\mu_i > \mu_j$, thus forcing the left-hand side of (3.1(h)) to be negative and the right-hand side of (3.1(h)) to be positive. This contradicts the assumption that $\phi_j(x)$ has no zeros on $(\sigma, \tilde{x})$.

We can now repeat this argument on $(\tilde{x}, \tilde{y})$ where $\tilde{y}$ is the next zero of $\phi_i(x)$. A similar argument integrating (3.1(f)) from $\tilde{x}$ to $\tilde{y}$ will show that $\phi_j(x)$ has a zero between $\tilde{x}$ and $\tilde{y}$. This argument can be repeated to show that there is a zero of $\phi_j(x)$ between any two consecutive zeros of $\phi_i(x)$.

In a similar fashion we see that if $\tilde{z}$ is the largest zero of $\phi_i(x)$ on (σ, ε), then $\phi_j(x)$ has a zero in $(\tilde{z}, \varepsilon)$. Thus ϕ_j has at least one more zero than ϕ_i $(j > i)$.

Noting that the ϕ_j's are polynomials of degree $\leqq L$ and thus have at most L zeros, we conclude that ϕ_j has at most j zeros. Furthermore we can show that since ϕ_0 has

no zeros, applying the argument above to the interval (σ, ε) (i.e., integrating (3.1(f)) from σ to ε) with $i=0$ and $j=1$ enables us to conclude that $\phi_1(x)$ has exactly one zero and that $\phi_j(x)$ has exactly j zeros.

We can now repeat the argument above on the interval (δ, σ) (see Lemma 2.1(v)), i.e., integrate (3.1(f)) from $\tilde{x}$ to σ $(\tilde{x}<\sigma)$, to show that if $\mu_i<\mu_j$, then ϕ_j has at least one more zero than ϕ_i on (δ, σ). Thus ϕ_j has $L-j$ zeros on (δ, σ).

The interlacing property of the zeros is immediate thus proving (c).

Part (d) follows from consideration of the matrix T. If T is oscillatory [8] then (d) follows immediately. A tridiagonal matrix is oscillatory if its super- and subdiagonal elements are positive and its successive principal minors are all positive [8]. Clearly in the Jacobi and Hermite cases the matrices T all have positive elements along the super- and subdiagonals but it need not be that the successive principal minors are all positive. Note, however, that if $TG=GT$, then also $(T+\lambda l)G=G(T+\lambda l)$ and $T+\lambda l$ has the same eigenvectors as T does, but has eigenvalues $\mu_0+\lambda>\mu_1+\lambda>\cdots>\mu_L+\lambda$. Since there are a finite number of principal minors of $T+\lambda l$, all of which are polynomials in λ with leading coefficient 1, we can choose λ sufficiently large so as to guarantee that the principal minors of $T+\lambda l$ are positive and so $T+\lambda l$ is oscillating. This assures (d) in these two cases.

In the Laguerre case the elements along the super- and subdiagonals of T are negative. We can replace T by $-T$ and carry out the same argument, only now the eigenvalues are $-\mu_L>-\mu_{L-1}>\cdots>-\mu_0$, and thus this reordering leads to $S^-(\vec{c}^{(i)})=L-i$.

We now consider (a). Note that $p(x)$ is an eigenfunction of $AF^{-1}BFA$ corresponding to $\lambda\neq 0$ if and only if

$$\int (\chi_{\mathscr{A}}-\lambda\chi_{\mathscr{C}})p(x)f(x)w(x)\,dx=0 \tag{3.2}$$

for all polynomials $f(x)$ of degree $\leqq L$. For if $p(x)$ is an eigenfunction of $AF^{-1}BFA$, then

$$\int_{\mathscr{A}} K_L(x,y)p(y)w(y)\,dy=\lambda p(x)=\lambda\int_{\mathscr{C}} K_L(x,y)p(y)w(y)\,dy.$$

The last equality is due to the fact that $p(y)$ is a polynomial of degree $\leqq L$ and $K_L(x,y)$ is a reproducing kernel for such polynomials. From this we see that

$$\int (\chi_{\mathscr{A}}-\lambda\chi_{\mathscr{C}})K_L(x,y)p(y)w(y)\,dy=0, \quad \text{or}$$

$$\sum_{i=0}^{L}\left(\int(\chi_{\mathscr{A}}-\lambda\chi_{\mathscr{C}})p_i(y)p(y)w(y)\,dy\right)p_i(x)=0$$

and so the coefficients of $p_0(x), p_1(x), \cdots, p_L(x)$ in the above sum must each be 0. Thus for any polynomial $f(x)$ of degree $\leqq L$ we have (3.2). Conversely if (3.2) holds for all polynomials of degree $\leqq L$, then we have

$$\begin{aligned}\int_{\mathscr{A}} K_L(x,y)p(y)w(y)\,dy&=\sum_{i=0}^{L}p_i(x)\left(\int_{\mathscr{A}}p_i(y)p(y)w(y)\,dy\right)\\&=\lambda\sum_{i=0}^{L}p_i(x)\int_{\mathscr{C}}p_i(y)p(y)w(y)\,dy\\&=\lambda\int_{\mathscr{C}}K_L(x,y)p(y)w(y)\,dy=\lambda p(x)\end{aligned}$$

and thus $p(x)$ is an eigenfunction.

Suppose now that $AF^{-1}BFA$ does not have simple spectrum. Then there exist eigenfunctions $f(x)$ and $g(x)$, linearly independent polynomials of degree $\leqq L$, corresponding to eigenvalue λ. Without loss of generality, we may assume one of these, say $g(x)$, has degree strictly less than L. Since $g(x)(\sigma - x)$ has degree $\leqq L$, by (3.2) we have

$$\int (\chi_{\mathscr{A}} - \lambda\chi_{\mathscr{C}})g^2(x)(\sigma - x)w(x)\,dx = 0. \tag{3.3}$$

But the left-hand side of (3.3) is in fact equal to

$$(1-\lambda)\int_{\mathscr{A}} g^2(x)(\sigma - x)w(x)\,dx - \lambda\int_{\sim\mathscr{A}} (\sigma - x)g^2(x)w(x)\,dx$$

and this is clearly less than 0, contradicting (3.3). Thus $AF^{-1}BFA$ has simple spectrum proving (a).

To prove (b), we claim that it suffices to show that the ordering in (b) holds for just one value of σ. For suppose that L is fixed and that we write $G(\sigma)$ for the matrix G of (1.2) in order to denote its dependence on σ. Since the elements of $G(\sigma)$ are analytic functions of σ and the eigenvalues $\lambda_j(\sigma)$ of $G(\sigma)$ are distinct, we know that the functions $\lambda_j(\sigma)$ are analytic functions of σ [12]. Now if the ordering in (b) holds for some particular σ and for some σ' the ordering does not hold, i.e., $\lambda_j(\sigma') > \lambda_{j-1}(\sigma')$, for some j, then by continuity considerations for some σ'' between σ and σ', $\lambda_{j-1}(\sigma'') = \lambda_j(\sigma'')$, contradicting (a). This shows the claim.

Before continuing the proof of (b), we make note of the following ideas from the theory of total positivity [13], [14], [15]. A kernel $K(x, y)$ is said to be totally positive on (α, β) [14] provided the Fredholm determinants

$$K\begin{pmatrix} x_0, \cdots, x_n \\ y_0, \cdots, y_n \end{pmatrix} = \det\,(K(x_i, y_j))_{i,j=0}^{n} \geqq 0 \tag{3.4}$$

for all $\alpha \leqq x_0 < x_1 < \cdots < x_n \leqq \beta$ and $\alpha \leqq y_0 < y_1 < \cdots < y_n \leqq \beta$ and $n = 0, 1, 2, \cdots$.

Gantmacher and Krein prove the following theorem [13, pp. 207–217].

THEOREM. *Let $K(x, y)$ be a totally positive continuous symmetric kernel on (α, β), and suppose that*

$$K\begin{pmatrix} x_0, \cdots, x_n \\ x_0, \cdots, x_n \end{pmatrix} > 0 \quad \textit{for all } \alpha < x_0 < x_1 < \cdots < x_n < \beta \textit{ and } n = 0, 1, 2, \cdots. \tag{3.5}$$

Then the integral equation

$$\int_{\alpha}^{\beta} K(x, y)f(y)w(y)\,dy = \lambda f(x)$$

has positive simple eigenvalues $\lambda_0 > \lambda_1 > \cdots > 0$. The corresponding eigenfunctions $\phi_0(x)$, $\phi_1(x), \cdots$ are such that ϕ_i has exactly i (simple) zeros in (α, β).

If we could show that for some σ, $K_L(x, y)$ satisfies the conditions of the theorem on (σ, ε) then we could conclude that, by the proof of (c), the ordering in (b) holds. In fact, a σ can be found so that $K_L(x, y)$ satisfies (3.4) and (3.5), but only for $n = 0, 1, \cdots, L$. With only minor modifications of the proof in [13], however, we see that if (3.5) holds only for $n = 0, 1, \cdots, L$ then the eigenfunctions $\phi_i(x)$ have exactly i zeros in (σ, ε) for $i = 0, 1, \cdots, L$.

In order to show that $K_L(x, y)$ satisfies (3.4) and (3.5) for $n = 0, 1, \cdots, L$ and some σ, we will make use of Karlin's notion of extended total positivity [11]. We call the kernel $K(x, y)$ extended totally positive of order r on (α, β) if

$$\tilde{K}^{(n)}(x, y) = \det\left(\frac{\partial^{i+j} K(x, y)}{\partial x^i \, \partial y^j}\right)_{i,j=0}^{n} > 0 \tag{3.6}$$

for $n = 0, 1, \cdots, r$ and for all $x, y \in (\alpha, \beta)$. A theorem of Karlin [11] states that if (3.6) holds then (3.4) holds with strict inequality for $n = 0, 1, \cdots, r$. Thus it suffices to show that there exists σ so that $\tilde{K}_L^{(n)}(x, y) > 0$ for $n = 0, 1, \cdots, L$ and $x, y \in (\sigma, \varepsilon)$.

Let

$$A^{(n)}(z) = \begin{bmatrix} p_0(z) & p_1(z) & \cdots & p_L(z) \\ p_0'(z) & p_1'(z) & \cdots & p_L'(z) \\ \vdots & \vdots & & \vdots \\ p_0^{(n)}(z) & p_1^{(n)}(z) & \cdots & p_L^{(n)}(z) \end{bmatrix}$$

$$= \begin{bmatrix} p_0(z) & p_1(z) & \cdots & p_n(z) & \cdots & p_L(z) \\ 0 & p_1'(z) & \cdots & p_n'(z) & \cdots & p_L'(z) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & p_n^{(n)}(z) & \cdots & p_L^{(n)}(z) \end{bmatrix}.$$

Then $\tilde{K}_L^{(n)}(x, y) = \det\left(A^{(n)}(y)(A^{(n)}(x))^T\right)$. We consider the cases of Jacobi, Hermite, and Laguerre polynomials separately.

(i) Jacobi polynomials. Noting that $A^{(L)}(z)$ is a square upper triangular matrix whose diagonal terms are nonzero constants, we have at once that

$$\tilde{K}_L^{(L)}(x, y) = \det\left(A^{(L)}(y)\right) \det\left(A^{(L)}(x)\right)^T = \left(\det\left(A^{(L)}(y)\right)\right)^2 > 0.$$

If we fix x then $\tilde{K}_L^{(n)}(x, x)$ is the determinant of the nonnegative definite matrix $A^{(n)}(x)(A^{(n)}(x)^T)$. For $n = L$ this matrix is positive definite ($\tilde{K}_L^{(L)}(x, x) > 0$) and thus the eigenvalues of this matrix are positive numbers

$$0 < \tau_0 \leqq \tau_1 \leqq \tau_2 \leqq \cdots \leqq \tau_L.$$

By the separation theorem for the eigenvalues of a symmetric matrix [10], the eigenvalues $\tau_0' \leqq \tau_1' \leqq \cdots \leqq \tau_{L-1}'$ of $A^{(L-1)}(x)(A^{(L-1)}(x))^T$ interlace with the τ_i's, and thus

$$0 < \tau_0 \leqq \tau_0' \leqq \tau_1 \leqq \tau_1' \cdots \leqq \tau_{L-1} \leqq \tau_{L-1}' \leqq \tau_L$$

and $\tilde{K}_L^{(L-1)}(x, x) = \tau_0' \cdots \tau_{L-1}' > 0$. Continuing in this fashion we see that $\tilde{K}_L^{(n)}(x, x) > 0$ for $n = 0, 1, \cdots, L$.

Since $\tilde{K}_L^{(n)}(x, y)$ is a continuous function of x and y for each n and since $\tilde{K}_L^{(n)}(1, 1) > 0$, we have that $\tilde{K}_L^{(n)}(x, y) > 0$ in a neighborhood of $(1, 1)$. In particular, if σ is sufficiently close to 1, $\tilde{K}_L^{(n)}(x, y) > 0$ for all $x, y \in (\sigma, 1)$ and $n = 0, 1, \cdots, L$.

(ii) Laguerre polynomials. Here we can basically repeat the argument for Jacobi polynomials, only this time on the set $\sim A = (0, \sigma)$, i.e., we consider the operator $\sim AF^{-1}BF \sim A$ of Lemma 2.1(v). The eigenvalues in this case are $1 > 1 - \lambda_L > 1 - \lambda_{L-1} > \cdots > 1 - \lambda_0 > 0$ with corresponding eigenvectors $\phi_L, \phi_{L-1}, \cdots, \phi_0$. The same D and T of Theorem 2.2 commute with $\sim AF^{-1}BF \sim A$ and $BF \sim AF^{-1}B$, respectively. By taking σ sufficiently close to 0 we can conclude that $\tilde{K}_L^{(n)}(x, y) > 0$ for all $x, y \in (0, \sigma)$, $n = 0, 1, \cdots, L$. Thus ϕ_i has $L - i$ zeros in $(0, \sigma)$ and thus i zeros in (σ, ∞), completing (b) of Theorem 3.1.

(iii) Hermite polynomials. In this case we will use the Cauchy-Binet formula [14]. This gives us that

$$\tilde{K}_L^{(n)}(x,y)=\sum_{1\leqq i_1<i_2<\cdots<i_n\leqq L} A^{(n)}(y)\begin{pmatrix}1&2&\cdots&n\\ i_1&i_2&\cdots&i_n\end{pmatrix}\cdot(A^{(n)}(x))^T\begin{pmatrix}i_1&i_2&\cdots&i_n\\ 1&2&\cdots&n\end{pmatrix} \tag{3.7}$$

where

$$B\begin{pmatrix}j_1&j_2&\cdots&j_n\\ i_1&i_2&\cdots&i_n\end{pmatrix}$$

is the determinant of the $n\times n$ matrix obtained by omitting all rows from B except $j_1,\cdots,j_n$ and all columns from B except $i_1,\cdots i_n$. Note that $A^{(n)}(y)\left(\begin{smallmatrix}1&2&\cdots&n\\ i_1&i_2&\cdots&i_n\end{smallmatrix}\right)$ is a polynomial in y. We will show that the leading coefficient of the polynomial is always positive. Thus if σ is sufficiently large, $A^{(n)}(y)\left(\begin{smallmatrix}1&2&\cdots&n\\ i_1&i_2&\cdots&i_n\end{smallmatrix}\right)>0$ for all $y>\sigma$ and all choices of $i_1<i_2<\cdots<i_n$ and $n=0,1,\cdots,L$. This in turn would enable us to conclude that every term in the sum (3.7) is positive for $x,y\in(\sigma,\infty)$ and thus that $\tilde{K}_L^{(n)}(x,y)>0$ on (σ,∞).

It remains only to show that the leading coefficient of

$$A^{(n)}(x)\begin{pmatrix}1&2&\cdots&n\\ i_1&i_2&\cdots&i_n\end{pmatrix}=\det\begin{bmatrix}p_{i_1}(x)&p_{i_2}(x)&\cdots&p_{i_n}(x)\\ p'_{i_1}(x)&p'_{i_2}(x)&\cdots&p'_{i_n}(x)\\ \vdots&\vdots&&\vdots\\ p_{i_1}^{(n)}(x)&p_2^{(n)}(x)&\cdots&p_{i_n}^{(n)}(x)\end{bmatrix}$$

is positive.

The $p_i(x)$'s are the orthonormalized Hermite polynomials and, since their leading coefficients are all positive, we may renormalize them to have leading coefficient 1 without affecting the sign of $A^{(n)}(x)\left(\begin{smallmatrix}1&\cdots&n\\ i_1&\cdots&i_n\end{smallmatrix}\right)$. Thus $A^{(n)}(x)\left(\begin{smallmatrix}1&2&\cdots&n\\ i_1&i_2&\cdots&i_n\end{smallmatrix}\right)$ is the determinant of

$$\begin{bmatrix}x^{i_1}+\cdots&x^{i_2}+\cdots&\cdots&x^{i_n}+\cdots\\ i_1x^{i_1-1}+\cdots&i_2x^{i_2-1}+\cdots&\cdots&i_nx^{i_n-1}+\cdots\\ 2!\binom{i_1}{2}x^{i_1-2}+\cdots&2!\binom{i_2}{2}x^{i_2-2}+\cdots&\cdots&2!\binom{i_n}{2}x^{i_n-2}+\cdots\\ \vdots&\vdots&&\vdots\\ (n-1)!\binom{i_1}{n-1}x^{i_1-(n-1)}&(n-1)!\binom{i_2}{n-1}x^{i_2-(n-1)}&\cdots&(n-1)!\binom{i_n}{n-1}x^{i_n-(n-1)}\end{bmatrix}$$

(here $\binom{i}{j}=0$ if $j>i$) and this determinant has x degree at most $i_1+i_2+\cdots+i_n-\binom{n}{2}$. Thus the coefficient of $x^{i_1+\cdots+i_n-\binom{n}{2}}$ is $(n-1)!(n-2)!\cdots 2!$ times the determinant of

$$\begin{bmatrix}1&1&1\\ \binom{i_1}{1}&\binom{i_2}{1}&\binom{i_n}{1}\\ \binom{i_1}{2}&\binom{i_2}{2}&\binom{i_n}{2}\\ \vdots&\vdots&\vdots\\ \binom{i_1}{n-1}&\binom{i_2}{n-1}&\binom{i_n}{n-1}\end{bmatrix}. \tag{3.8}$$

That determinants of matrices of the form (3.8) are greater than or equal to 0 follows immediately from the fact that these determinants are minors of the transpose of the matrix $(B)_{ij} = \binom{i}{j}$, $i, j = 0, \cdots, N$, and all minors of B are known to be greater than or equal to 0 [15, p. 50]. Suppose now that the determinant of (3.8) is 0. Then the n rows of the matrix (3.8) must be linearly dependent and thus there exist constants $c_0, c_1, \cdots, c_{n-1}$, not all 0, so that

$$c_0 + c_1 x + c_2 \frac{(x)(x-1)}{2!} + \cdots + c_{n-1} \frac{(x)(x-1)\cdots(x-(n-2))}{(n-1)!} \tag{3.9}$$

has n distinct roots, $i_1, i_2, \cdots, i_n$. But the polynomial in (3.9) has degree $n-1$, a contradiction. Thus the determinant of (3.8) is always positive. This completes (b) of Theorem 3.1.

4. Comments. In [16] Gilbert and Slepian looked at the operator $AF^{-1}BFA$ for Legendre polynomials with $\mathcal{A} = (-\sigma, \sigma)$, where $0 < \sigma < 1$. They sought a commuting differential operator. The results in this case were more complex and simplicity of spectrum did not follow. The general techniques they employed in studying this problem, however, form the basis of the argument in the proof of Theorem 3.1(a). We further note that this argument applies quite generally to show that $AF^{-1}BFA$ has simple spectrum away from 0 (regardless of whether or not the underlying polynomials are classical ones) as long as $\mathcal{A} = (\sigma, \varepsilon)$.

The fact that the matrix T is oscillating as noted in the proof of Theorem 3.1(d) means that much stronger results can be stated about the eigenvectors $\vec{c}^{(i)}$. See, for example, [8] and [15].

In the case of the standard Fourier transform discussed in [1], applications have arisen based on the oscillation properties of the eigenfunctions [17]. Some related applications for the polynomial cases will be discussed elsewhere.

REFERENCES

[1] D. SLEPIAN AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: I, Bell System Tech. J., 40 (1961), pp. 43-64.

[2] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: II, Bell System Tech. J., 40 (1961), pp. 65-84.

[3] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty*: III, Bell System Tech. J., 41 (1962), pp. 1295-1336.

[4] F. A. GRÜNBAUM, L. LONGHI, AND M. PERLSTADT, *Differential operators commuting with finite convolution integral operators: Some nonabelian examples*, SIAM J. Appl. Math, 42 (1982), pp. 941-955.

[5] F. A. GRÜNBAUM, *A new property of reproducing kernels for classical orthogonal polynomials*, J. Math. Anal. Appl., 95 (1983), pp. 491-500.

[6] M. PERLSTADT, *Polynomial analogues of prolate spheroidal wave functions and uncertainty*, SIAM J. Math. Anal., 17 (1986), pp. 240-248.

[7] P. M. MORSE AND H. FESHBACH, *Methods of Theoretical Physics*, McGraw-Hill, New York, 1953.

[8] F. R. GANTMACHER, *The Theory of Matrices, Vol.* II, Chelsea, New York, 1971.

[9] W. MAGNUS, F. OBERHETTINGER, AND R. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, New York, 1966.

[10] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

[11] S. KARLIN, *Total positivity and convexity preserving transformations*, Proc. Symposia in Pure Mathematics, Vol. VII, Convexity, Amer. Math. Soc., Providence, RI, 1963.

[12] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, Orlando, 1985.

[13] F. R. GANTMACHER AND M. G. KREIN, *Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme*, Akademie-Verlag, Berlin, 1960.

[14] S. KARLIN, *Total Positivity Vol.* I, Stanford University Press, Stanford, CA, 1968.

[15] A. PINKUS, *n-Widths in Approximation Theory*, Springer-Verlag, New York, 1985.
[16] E. N. GILBERT AND D. SLEPIAN, *Doubly concentrated orthogonal polynomials*, SIAM J. Math. Anal., 8 (1977), pp. 290–319.
[17] A. A. MELKMAN, *n-Widths and optimal interpolation of time- and band-limited functions*, in Optimal Estimation and Approximation Theory, C. A. Micchelli and T. J. Rivlin, eds., Plenum, New York, 1977, pp. 55–68.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

REGULARITY THROUGH APPROXIMATION FOR SCALAR CONSERVATION LAWS*

BRADLEY J. LUCIER†

Abstract. In this paper it is shown that recent approximation results for scalar conservation laws in one space dimension imply that solutions of these equations with smooth, convex fluxes have more regularity than previously believed. Regularity is measured in spaces determined by quasinorms related to the solution's approximation properties in $L^1(\mathbf{R})$ by discontinuous, piecewise linear functions. Using a previous characterization of these approximation spaces in terms of Besov spaces, it is shown that there is a one-parameter family of Besov spaces that are invariant under the differential equation. An intriguing feature of this investigation is that regularity is measured quite naturally in smoothness classes that are not locally convex—they are similar to L^p spaces for $0 < p < 1$. Extensions to Hamilton-Jacobi equations are mentioned.

Key words. regularity, approximation, Besov spaces, conservation laws

AMS(MOS) subject classifications. 35L65, 35D10, 41A25, 46E35

1. Introduction.

It is well known that discontinuities may form in the solution $u(x,t)$ of the hyperbolic conservation law

$$\text{(C)} \qquad \begin{aligned} &u_t + f(u)_x = 0, \qquad x \in \mathbf{R}, \quad t > 0, \\ &u(x,0) = u_0(x), \qquad x \in \mathbf{R}, \end{aligned}$$

even if the flux f and the initial data u_0 are smooth. (In gas dynamics these discontinuities represent shocks.) Hence, classical solutions of (C) do not generally exist. Weak solutions of (C) are not unique, but both existence and uniqueness of weak solutions that satisfy an auxiliary "entropy" condition were shown by Vol'pert [22] and Kružkov [15]. The regularity of these weak solutions is the topic of this paper.

There have been two different approaches to studying the regularity of solutions of hyperbolic conservation laws of one or more independent variables. Both approaches are "structural" in that they describe properties of the solution without quantifying a norm, seminorm, or quasinorm that says, for example, that one function is twice as smooth as another. The first approach is to show that "generic" solutions of the scalar equation (C) with C^∞ initial data are piecewise C^∞. This approach has been followed, for example, by Schaeffer [21], Guckenheimer [13], and Dafermos [5], [6]. A typical result is that except for a set of first Baire category in the Schwartz class $\mathcal{S}$, initial data in $\mathcal{S}$ results in piecewise C^∞ solutions $u(x,t)$. (Various assumptions are made on the flux f, typically that it is convex or has isolated points of inflection.)

The second, more measure-theoretic, approach is to show that the set of singularities of a solution $u(x,t)$ is more restricted than those of an arbitrary function of bounded variation in $\mathbf{R} \times \mathbf{R}^+$. Consider the following definitions. If $u(x,t) \in \mathrm{BV}(\mathbf{R}^2)$, then it is known [11], [22] that for every point (x,t) outside of a set of one-dimensional Hausdorff measure zero (called the set of singular points), there exist numbers u^+ and

*Received by the editors July 8, 1987; accepted for publication December 18, 1987. This work was supported in part by the National Science Foundation under grant DMS-8403219 and by the Institute for Mathematics and its Applications with funds provided by the National Science Foundation.

†Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

u^- and a direction $\nu \in \mathbb{R}^2$ such that

$$\lim_{r\to 0} \frac{1}{r^2} \iint_{\{(y,\tau)\cdot(\pm\nu)\geq 0\}\cap B((x,t),r)} |u(y,\tau) - u^{\pm}|\, dy\, d\tau = 0.$$

If $u^+ = u^-$, then (x,t) is a point of approximate continuity; if $u^+ \neq u^-$, then (x,t) is a point of approximate discontinuity (a jump point). Furthermore, the set of regular points consisting of the jump points is at most a countable union of rectifiable sets of dimension $n-1$. DiPerna [9], [10] showed for genuinely nonlinear systems of two equations that the singular set of any solution u constructed by the random choice method of Glimm [12] is in fact at most countable, and that at each regular point of u, u has true one-sided limits that satisfy the Rankine-Hugoniot conditions. Furthermore, the shock set of u has "nice" structure.

In a similar vein, Oleinik [18] has shown that if f is convex, then u is continuous except on the union of a countable set of Lipschitz continuous curves (shocks). Dafermos [6] and Liu [16] establish similar results.

Rather than considering structural properties of solutions, either of the solution values (e.g., smoothness) or solution singularities (e.g., shocks), I consider smoothness in certain approximation spaces that are closely related to Besov spaces. I show that if u_0 is in one of these approximation spaces, then $u(\cdot\,,t)$ is in the same space for all later time if f is convex and smooth enough. (Of course, the results in this paper also hold if f is concave.) These function spaces are not Banach spaces, and are not even locally convex topological vector spaces, but they are composed of functions that are, in some sense, smoother than arbitrary functions in BV, or even arbitrary piecewise C^∞ functions (see §6). In §2 I rationalize this approach by arguing that $\mathrm{BV}(\mathbb{R})$ is the wrong space in which to measure smoothness, precisely because it is a locally convex topological vector space. The convexity of the "unit ball" of $\mathrm{BV}(\mathbb{R})$ allows only coarse measurement of the smoothness of functions that are discontinuous.

In this paper I consider a function smooth if it can be approximated well in L^1 by possibly discontinuous, piecewise linear functions with free knots—the better the approximation, the better the smoothness. This notion is developed in §3, in which I recount certain results of DeVore and Popov [7], [8], based on work by Petrushev [19], [20], that characterize the approximation spaces used here.

In §4, results from [17] are used to show that solutions of (C) that are initially in $\mathrm{BV}(\mathbb{R})$ preserve whatever smoothness is obtained by the initial data in the sense given in §3. In particular, it is shown that there is a one-parameter family of Besov spaces that are invariant under the action of (C) provided the initial data is of bounded variation. In §5, I point out that this is indeed a new result, because $\mathrm{BV}(\mathbb{R})$ is not contained in the approximation spaces when the order of smoothness is greater than one.

In §6, I show that there is, in general, no smoothing by the solution operator of (C) in the approximation spaces considered here. This result follows from the partial reversibility in time of the equation (C). The question arises because it is known that if f is uniformly convex, then initial data in $L^1(\mathbb{R})$ generate solutions that have locally bounded variation for all positive time, so there is some smoothing action from $L^1(\mathbb{R})$ to $\mathrm{BV}(\mathbb{R})$.

These ideas also have applications to regularity of solutions of Hamilton-Jacobi equations based on approximation properties in L^∞; this will be explored in a later paper. However, in §7 I present one result that follows immediately from the results in §4.

2. Why nonconvex spaces are natural. I begin with a specific example. Let u_0 be the characteristic function of $[0,1]$ and let $f(u) = u$. Then the solution $u(\cdot, t)$ of (C) is the characteristic function of $[t, 1+t]$. Except for the two jumps at the points t and $1+t$, $u(\cdot, t)$ is a very smooth function of x for every t. If the space that one uses to measure regularity allows any jumps at all (which it must, because solutions of (C) can develop jumps even for smooth data when f is nonlinear), then $u(\cdot, t)$ must be a relatively smooth function in that space.

Consider the inclusion of the functions $u(\cdot, t)$ in $\mathrm{BV}(\mathbb{R})$, or in fact in any normed or seminormed space whose unit ball is convex, and define the smoothness of $u(\cdot, t)$ to be its norm in this space. The solution $u(\cdot, t)$, being a translation of u_0, must have the same smoothness as u_0. (Of course, I am assuming that the norm or seminorm is translation invariant.) This implies that any convex linear combination of $u(\cdot, t)$ (for $0 \le t \le 1$, say) will also have the same smoothness, because the unit ball of $\mathrm{BV}(\mathbb{R})$ is convex.

It is easily seen that convex linear combinations of $u(\cdot, t)$ can approximate arbitrarily well in $L^1([0,1])$ any monotone function that takes the values 0 at 0 and 1 at 1. But, as is shown in §5 in a particular technical sense, an arbitrary monotone function is very rough, in that one can say very little, a priori, about the size and distribution of discontinuities in the interval $[0,1]$, for example, except that the sum of the jumps is bounded.

Thus, the convex hull of the solutions $u(\cdot, t)$ of (C) for our chosen u_0 contains functions that are quite rough. It is shown in §4 that *these rough functions cannot arise as solutions to* (C) *if* u_0 *and* f *are smooth enough.* It is in this sense that one discards information when one concludes that the solution of (C) at any particular time t has exactly the same smoothness as all functions in the convex hull of $u(\cdot, t)$ for $t > 0$. I conclude that it is better to measure the smoothness of solutions of (C) in spaces whose "unit balls" are not convex.

3. Approximation spaces and Besov spaces. Smoothness will be defined by how well a function can be approximated by piecewise polynomials with free knots. This section summarizes results in [7] and [8], which are given as general references for this section.

Consider the approximation of functions in $L^p(I)$ for $0 < p < \infty$ and a finite interval $I \subset \mathbb{R}$. For any $f \in L^p(I)$ and any positive integers r and N, let

$$E_N^r(f, L^p(I)) = \inf \|f - \phi\|_{L^p(I)},$$

where the infimum is taken over all discontinuous, piecewise polynomial functions ϕ defined on I of degree less than r with $N-1$ free interior knots. In other words, for each function f and number N one picks the best set of knots to minimize $\|f - \phi\|_{L^p(I)}$.

For each positive number α choose an integer $r > \alpha$. For any $q \in (0, \infty]$, define $\mathcal{A}_q^\alpha(L^p(I))$ to be the set of functions for which

$$\|f\|_{\mathcal{A}_q^\alpha(L^p(I))} = \|f\|_{L^p(I)} + \left(\sum_{N=1}^{\infty} [N^\alpha E_N^r(f, L^p(I))]^q N^{-1} \right)^{1/q} < \infty.$$

(In this and all later cases, make the usual modification when $q = \infty$.) It can be shown that all values of r greater than α specify the same space. Note that α is the primary determinant of smoothness: If $\alpha_1 > \alpha_2$, then no matter the value of

q_1 and q_2, $\mathcal{A}_{q_1}^{\alpha_1}(L^p(I)) \subset \mathcal{A}_{q_2}^{\alpha_2}(L^p(I))$. However, if $\alpha_1 = \alpha_2 = \alpha$ and $q_1 > q_2$, then $\mathcal{A}_{q_1}^{\alpha}(L^p(I)) \supset \mathcal{A}_{q_2}^{\alpha}(L^p(I))$.

The spaces $\mathcal{A}_q^\alpha(L^p(I))$ are not as strange as they might seem. If one denotes by $\mathcal{A}_q^\alpha(L^p(I), \text{free})$ the spaces described above, and by $\mathcal{A}_q^\alpha(L^p(I), \text{uniform})$ the similar spaces defined by considering approximation using only uniform knot sequences, then the space $\mathcal{A}_q^\alpha(L^p(I), \text{uniform})$ is the Besov space $B_q^\alpha(L^p(I))$ given below (cf. [7]). Also, if α is not an integer and $1 \le p < \infty$, then $\mathcal{A}_p^\alpha(L^p(I), \text{uniform})$ is the Sobolev space $W^{\alpha,p}(I)$ (cf. [1, p. 223]). Thus, there is a strong connection between approximation spaces and more classical function spaces.

$\mathcal{A}_q^\alpha(L^p(I))$ can be characterized as the interpolation space of $L^p(I)$ and certain Besov spaces using the real method of interpolation. For $\alpha \in (0, \infty)$ and $q \in (0, \infty]$, define the Besov space $B_q^\alpha(L^p(I))$ as follows. Pick any integer $r > \alpha$; let $\Delta^r(f,h)(x)$ be the rth forward difference of f at x with interval h;[1] and let $I_h = \{x \in I \mid x + rh \in I\}$. Define

$$w_r(f,t)_{L^p(I)} = \sup_{|h|<t} \|\Delta^r(f,h)\|_{L^p(I_h)}.$$

The Besov space $B_q^\alpha(L^p(I))$ is defined to be the set of functions f for which

$$|f|_{B_q^\alpha(L^p(I))} \equiv \left(\int_0^\infty [t^{-\alpha} w_r(f,t)_{L^p(I)}]^q \, dt/t \right)^{1/q}$$

is finite. Set $\|f\|_{B_q^\alpha(L^p(I))} = \|f\|_{L^p(I)} + |f|_{B_q^\alpha(L^p(I))}$. I specifically require the case when p and q are less than one.

The real method of interpolation using K-functionals can be described as follows. For any two spaces X_0 and X_1 contained in some larger space X, define the following functional for all f in $X_0 + X_1$:

$$K(f,t,X_0,X_1) = \inf_{f=f_0+f_1} \{\|f_0\|_{X_0} + t\|f_1\|_{X_1}\},$$

where $f_0 \in X_0$ and $f_1 \in X_1$. The new space $X_{\theta,q} = (X_0, X_1)_{\theta,q}$ $(0 < \theta < 1$, $0 < q \le \infty)$ consists of functions f for which

$$\|f\|_{X_{\theta,q}} = \|f\|_{X_0+X_1} + \left(\int_0^\infty [t^{-\theta} K(f,t,X_0,X_1)]^q \, dt/t \right)^{1/q} < \infty,$$

where $\|f\|_{X_0+X_1} = K(f,1,X_0,X_1)$. Using results of Petrushev [19], [20], the following theorem is proved in [8].

THEOREM 3.1 (DeVore and Popov). *When $0 < p < \infty$, $0 < q \le \infty$, and $0 < \alpha < \beta$, define $\sigma = 1/(\beta + 1/p)$. Then*

$$\mathcal{A}_q^\alpha(L^p(I)) = (L^p(I), B_\sigma^\beta(L^\sigma(I)))_{\alpha/\beta,q},$$

and if $q = 1/(\alpha + 1/p)$,

$$\mathcal{A}_q^\alpha(L^p(I)) = B_q^\alpha(L^q(I)).$$

Thus, there is a two-parameter family of spaces $\mathcal{A}_q^\alpha(L^p(I))$ that are Besov spaces, albeit with q possibly less than 1.

[1] Set $\Delta^0(f,h)(x) = f(x)$ and $\Delta^r(f,h)(x) = \Delta^{r-1}(f,h)(x+h) - \Delta^{r-1}(f,h)(x)$.

Although there is not now an exact characterization of all the spaces $\mathcal{A}_q^\alpha(L^p(I))$ in terms of Besov or other spaces, the above theorem allows one to make rather precise statements about inclusions of these spaces in Besov spaces. For example, if $0 < q < 1/(\alpha + 1/p)$, $\beta > \alpha$, $\tilde{\beta} = 1/(\beta + 1/p)$, and $\tilde{\alpha} = 1/(\alpha + 1/p)$, then

$$B_{\tilde{\beta}}^{\beta}(L^{\tilde{\beta}}(I)) = \mathcal{A}_{\tilde{\beta}}^{\beta}(L^p(I)) \subset \mathcal{A}_q^\alpha(L^p(I)) \subset \mathcal{A}_{\tilde{\alpha}}^{\alpha}(L^p(I)) = B_{\tilde{\alpha}}^{\alpha}(L^{\tilde{\alpha}}(I)).$$

There is an atomic decomposition for functions in $B_q^\alpha(L^p(I))$; see [7] for details.

4. Regularity for scalar conservation laws. I modify several results in [17] to prove Theorem 4.2, which is the major result of this paper. The definitions from §3 will be used, assuming always now that $L^p = L^1$. First, I prove the following lemma.

LEMMA 4.1. *There is a constant C_1 such that for all u_0 in $\mathrm{BV}(\mathbb{R})$ with support in $[0,1]$ and for any N, the best $L^1([0,1])$, discontinuous, piecewise linear approximation with $N-1$ free interior knots v_0 to u_0 satisfies $|v_0|_{\mathrm{BV}(\mathbb{R})} \le C_1 |u_0|_{\mathrm{BV}(\mathbb{R})}$.*

Proof. Let $\{\tau_i\}_{i=0}^N$, with $\tau_0 = 0$ and $\tau_N = 1$, be the ordered set of knots of v_0. Consider now only one interval $I_i = (\tau_i, \tau_{i+1})$; let $\Delta\tau = \tau_{i+1} - \tau_i$, and let $\overline{u} = \sup_{x \in I_i} u_0(x)$, $\underline{u} = \inf_{x \in I_i} u_0(x)$, and $\Delta u = \overline{u} - \underline{u}$. Let s be the slope of v_0 in I_i.

If $|s|\Delta\tau > \Delta u$, then it is easily calculated that the $L^1(I_i)$ difference between u_0 and v_0 is at least

$$\frac{|s|}{4}\left(\Delta\tau - \frac{\Delta u}{s}\right)^2,$$

which is simply the area of the set of points that are greater than $\overline{u}$ but less than v_0 plus those points that are less than $\underline{u}$ but greater than v_0. If $|s| > 2(1+\sqrt{3})\Delta u/\Delta\tau$, then this error is greater than the error of the constant approximation $v_0 \equiv (\overline{u} + \underline{u})/2 = \tilde{u}$, so one must conclude that $|s| \le 2(1+\sqrt{3})\Delta u/\Delta\tau$. Thus

$$\begin{aligned}\mathrm{Var}_{I_i} v_0 &= |s|\Delta\tau \\ &\le 2(1+\sqrt{3})\Delta u \\ &\le 2(1+\sqrt{3})\mathrm{Var}_{I_i} u_0.\end{aligned}$$

So $\sum_i \mathrm{Var}_{I_i} v_0 \le 2(1+\sqrt{3})|u_0|_{\mathrm{BV}(\mathbb{R})}$.

Consider now the jump $|v_0(\tau_i^+) - v_0(\tau_i^-)|$. Subscript the quantities s, $\overline{u}$, $\underline{u}$, $\tilde{u}$, $\Delta\tau$, and Δu to indicate the interval I_i to which they pertain. Without loss of generality, assume that $s_{i-1} > 0$ and $s_i > 0$. Then $v_0(\tau_i^-) \le \tilde{u}_{i-1} + (1+\sqrt{3})\Delta u_{i-1}$ and $v_0(\tau_i^+) \ge \tilde{u}_i - (1+\sqrt{3})\Delta u_i$. So

$$\begin{aligned}|v_0(\tau_i^+) - v_0(\tau_i^-)| &\le |\tilde{u}_{i-1} - \tilde{u}_i| + (1+\sqrt{3})(|\Delta u_i| + |\Delta u_i|) \\ &\le \mathrm{Var}_{I_{i-1}\cup I_i} u_0 + (1+\sqrt{3})\mathrm{Var}_{I_{i-1}\cup I_i} u_0 \\ &\le (2+\sqrt{3})\mathrm{Var}_{I_{i-1}\cup I_i} u_0.\end{aligned}$$

So, $\sum_i |v_0(\tau_i^+) - v_0(\tau_i^-)| \le (4+2\sqrt{3})|u_0|_{\mathrm{BV}(\mathbb{R})}$. Adding these two constants will give the required value of C_1. □

The previous argument can be extended to show that the ranges of u_0 and all best piecewise linear approximations v_0 are uniformly bounded and contained in some interval, here denoted by Ω.

THEOREM 4.1 (Approximation). *Let $u_0 \in \mathrm{BV}(\mathbb{R})$ have support in the interval $I = [0,1]$. Assume that $f'' \geq 0$ and that f' and f''' are bounded on Ω. Then $u(\cdot,t)$ has support in $I_t = [\inf_{\xi\in\Omega} f'(\xi)t, 1 + \sup_{\xi\in\Omega} f'(\xi)t]$, $|u(\cdot,t)|_{\mathrm{BV}(\mathbb{R})} \leq |u_0|_{\mathrm{BV}(\mathbb{R})}$, and for any $N \geq 1$,*

$$E^2_{F(N)}(u(\cdot,t), L^1(I_t)) \leq E^2_N(u_0, L^1(I)) + \frac{t}{4N^2}|u_0|_{\mathrm{BV}(\mathbb{R})}\|f'''\|_{L^\infty}, \tag{4.1}$$

where $F(N) = \lfloor (C_1|u_0|_{\mathrm{BV}(\mathbb{R})} + 4)N + 4 \rfloor$ and C_1 is given in Lemma 4.1.

Proof. I will not discuss the first two conclusions of the theorem, which are classical. The proof of the third part models very closely the proofs of Theorems 3 and 4 in [17]. However, for the sake of completeness, I will recall the major parts of that paper.

Let v_0 be the best $L^1(I)$, discontinuous, piecewise linear approximation with $N-1$ free knots to u_0. Then, as shown in Lemma 4.1, $|v_0|_{\mathrm{BV}(\mathbb{R})} \leq C_1|u_0|_{\mathrm{BV}(\mathbb{R})}$. Consider the C^1, piecewise quadratic function g with knots at the points j/N, $j \in \mathbb{Z}$, that is defined by: $g'(j/N) = f'(j/N)$ and $g(0) = f(0)$. In [17] I constructed an explicit solution to the perturbed problem

$$\text{(P)} \qquad \begin{aligned} &v_t + g(v)_x = 0, && x \in \mathbb{R}, \quad t > 0, \\ &v(x,0) = v_0(x), && x \in \mathbb{R}, \end{aligned}$$

provided that one augments the knots of v_0 by putting a new knot at each isolated point x for which $v_0(x) = j/N$ for some j. (If v_0 is discontinuous at x, and there are k values of j such that $\min(v_0(x^-), v_0(x^+)) < j/N < \max(v_0(x^-), v_0(x^+))$, then add k knots at the point x.) Although these knots are not needed for the definition of v_0, the solution $v(\cdot,t)$ of (P) may develop discontinuities in its first derivative ("kinks") at these new knots for positive times.

The new knots number no more than $(2 + |v_0|_{\mathrm{BV}(\mathbb{R})})N + 1$, by the following argument. Let the original knots of v_0 be $\tau_0 = 0 < \tau_1 < \cdots < \tau_N = 1$, let $\sigma_{2i} = \sigma_{2i+1} = \tau_i$ for $i = 0, \ldots, N$, and consider the B-spline basis for v_0 with the knots $\{\sigma_i\}$. (See de Boor [2, Chap. 9] for this construction.) For each i, let k_i denote the number of original intervals (σ_j, σ_{j+1}) that had i new knots added. The value of $\sum_i ik_i$ is to be bounded. Now, $\sum_i k_i = 2N + 1$. But if i points are added in an original interval (σ_j, σ_{j+1}), the variation of v_0 in that interval must be at least $(i-1)/N$, so $\sum_i (i-1)k_i/N \leq |v_0|_{\mathrm{BV}(\mathbb{R})}$, or $\sum_i (i-1)k_i \leq N|v_0|_{\mathrm{BV}(\mathbb{R})}$. Adding these two known inequalities shows that $\sum_i ik_i \leq (2 + |v_0|_{\mathrm{BV}(\mathbb{R})})N + 1$, as claimed. Thus, the total number of knots in v_0 (counting the points σ_i and the new points, all of which may travel along different characteristics for t positive) is bounded by $(4 + C_1|u_0|_{\mathrm{BV}(\mathbb{R})})N + 3$.

It is shown in [17] that $v(\cdot,t)$ is piecewise linear for all time and that the number of knots decreases monotonically, because f'' is nonnegative. Theorem 3 of [17] shows that

$$\|u(\cdot,t) - v(\cdot,t)\|_{L^1(\mathbb{R})} \leq \|u_0 - v_0\|_{L^1(\mathbb{R})} + t\|f' - g'\|_{L^\infty}|u_0|_{\mathrm{BV}(\mathbb{R})}.$$

Because of the way g is constructed, $\|f' - g'\|_{L^\infty} \leq \|f'''\|_{L^\infty}/(4N^2)$, so (4.1) follows immediately. □

The proof can be easily modified to cover the case where $f \in C^1$ and is piecewise C^3 on intervals I_j with $\inf |I_j|$ positive.

Theorem 4.1 can be used to prove the following main result of this paper.

THEOREM 4.2 (Regularity). *Assume that there is an* $\alpha \in (0,2)$ *and a* $q \in (0,\infty]$ *such that* u_0 *has support in* $[0,1]$ *and* $u_0 \in \mathrm{BV}(\mathbb{R}) \cap \mathcal{A}_q^\alpha(L^1([0,1]))$. *Assume that* $f'' \geq 0$ *and that* f' *and* f''' *are bounded on* Ω. *Then* $u(\cdot,t)$ *has support in* $I_t = [\inf_{\xi\in\Omega} f'(\xi)t, 1 + \sup_{\xi\in\Omega} f'(\xi)t]$ *and* $u(\cdot,t) \in \mathrm{BV}(\mathbb{R}) \cap \mathcal{A}_q^\alpha(L^1(I_t))$.

Proof. Inequality (4.1) shows that the error in approximation (by piecewise linear functions) of $u(\cdot,t)$ is no more than the error in approximation of u_0 plus something of $O(N^{-2})$, and that the number of knots remains $O(N)$ for all later times. This is sufficient to show that $u(\cdot,t) \in \mathcal{A}_q^\alpha(L^1(I_t))$ if $\alpha < 2$. □

By combining Theorem 4.2 and the characterization of the spaces $\mathcal{A}_q^\alpha(L^1(I_t))$ in terms of Besov spaces, the following corollary is obtained.

COROLLARY 4.1. *Let* $0 < \alpha < 2$, *and set* $q = 1/(\alpha+1)$. *If* u_0 *has support in* $I = [0,1]$ *and* $u_0 \in \mathrm{BV}(\mathbb{R}) \cap B_q^\alpha(L^q(I))$, $f'' \geq 0$ *and* f' *and* f''' *are bounded on* Ω, *then* $u(\cdot,t)$ *has support in* $I_t = [\inf_{\xi\in\Omega} f'(\xi)t, 1 + \sup_{\xi\in\Omega} f'(\xi)t]$ *and* $u(\cdot,t) \in \mathrm{BV}(\mathbb{R}) \cap B_q^\alpha(L^q(I_t))$.

Thus, there is a one-parameter family of Besov spaces that are invariant under the action of the semigroup S_t that takes u_0 to $u(\cdot,t)$.

Theorem 4.2 and Corollary 4.1 are of interest only when α is greater than one, because any function in $\mathrm{BV}([0,1])$ can be approximated to within $O(N^{-1})$ in $L^1([0,1])$ by piecewise constant functions with $N-1$ uniformly spaced knots, so $\mathrm{BV}([0,1]) \subset \mathcal{A}_q^\alpha(L^1([0,1]))$ when $0 < \alpha < 1$, or when $\alpha = 1$ and $q = \infty$.

5. Approximation spaces and BV. In this section I give examples of the known fact that $\mathcal{A} \equiv \mathcal{A}_q^\alpha(L^1([0,1])) \not\subseteq \mathrm{BV}([0,1])$ and, if α is greater than one, $\mathrm{BV}([0,1]) \not\subseteq \mathcal{A}$.

First, I present an increasing function ϕ in $\mathrm{BV}([0,1])$ but not in $\mathcal{A}$ for any α greater than one. The function ϕ will take the value 0 to the left of 0 and will take the value $\pi^2/6$ to the right of 1. Its definition is as follows.

The jumps of ϕ will be at the points $p/2^k$ for p an odd integer between 1 and $2^k - 1$ with k a positive integer. For each k, the size of the jump at the point $p/2^k$ will be $1/(k^2 2^{k-1})$. Between the jumps, ϕ will be constant, so that if one arbitrarily defines ϕ to be right continuous, ϕ is given by the formula

$$\phi(x) = \sum_{\substack{p/2^k < x \\ k>0,\ 0<p<2^k,\ p \text{ odd}}} \frac{1}{k^2 2^{k-1}}.$$

Because for each k there are 2^{k-1} odd integers p between 0 and 2^k, $\phi(1)$, which is the sum of the jumps, is indeed $\pi^2/6$. Figure 1 is a graph of $\phi(x)/\phi(1)$.

Now consider the approximation of this function ϕ by possibly discontinuous linear functions with $2^M - 1$ interior knots for some positive M. Because ϕ behaves in exactly the same way on each interval $(j/2^M, (j+1)/2^M)$, it can be shown that the optimal placement of knots will be at the points $j/2^M$ for $0 < j < 2^M$. Because there is a jump of height $1/((M+1)^2 2^M)$ in the center of each interval $(j/2^M, (j+1)/2^M)$ and the width of the interval is $1/2^M$, the best linear approximation on this interval will have error greater than $C/((M+1)^2 2^{2M})$. Summing these errors over the 2^M intervals gives a global error of greater than $C/((M+1)^2 2^M)$, or, if one sets $N = 2^M$, $C/(\log^2(N)N)$. This quantity is asymptotically greater than C/N^α for any α greater than one, so ϕ is not in $\mathcal{A}$. Thus, one can conclude that if the conditions of Theorem

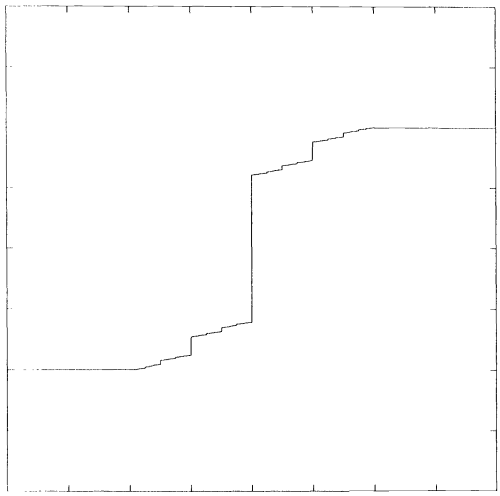


FIG. 1. *A function in* $\mathrm{BV}([0,1])$ *but not in* $\mathcal{A}_q^\alpha(L^1([0,1]))$ *for any* $\alpha > 1$.

4.2 hold with $\alpha > 1$, then this function ϕ cannot be the solution $u(\cdot, t)$ of (C) for any positive time t.

It is perhaps simpler to construct a function ϕ in $\mathcal{A}$ that is not of bounded variation. For x between 0 and 1, define

$$\phi(x) = \begin{cases} 0 & \text{for } 2^{-N} \le x < 1.5 \cdot 2^{-N},\ N > 0, \\ 1 & \text{for } 1.5 \cdot 2^{-N} \le x < 2^{-N+1},\ N > 0, \end{cases}$$

with $\phi(x) = 0$ for other values of x. (See Fig. 2.) It is clear that $\phi(x)$ can be approximated exactly by a piecewise *constant* function ψ with $2N$ knots for $2^{-N} < x < 1$. By setting $\psi(x) = 0$ for x greater than 0 and less than 2^{-N}, one obtains a global error in $L^1(\mathbb{R})$ of less than 2^{-N} with $O(N)$ knots. In other words, this ϕ can be approximated exponentially well by piecewise constant functions, and hence is in $\mathcal{A}$ for any values of α and q, yet ϕ is not of bounded variation.

Thus, the class $\mathcal{A}$ says little about the size of the jumps by themselves, but more about the combination of the size and *distribution* of the jumps in the functions. The example of a function of bounded variation but not in $\mathcal{A}$ had its jumps distributed uniformly in the interval $[0,1]$, thereby inhibiting good approximation by piecewise linear functions. In contrast, the example of a function in $\mathcal{A}$ but not of bounded variation had its jumps concentrated in a very small region. One may conclude intuitively that solutions of (C) that satisfy the hypotheses of Theorem 4.2 may be rough, but they are rough only in very small regions. This intuition is quantified in the atomic decomposition formula given in [7] for functions in Besov spaces.

6. Lack of smoothing. There is, in general, no smoothing in the spaces $\mathcal{A}_q^\alpha(L^1(I_t))$ for solutions of (C) as t progresses, even if the flux f is uniformly convex. This follows because of the partial reversibility of (C), as described below.

Define initial data u_0 as follows: Let $u_0(x)$ be zero for x less than 0 and greater than R, and constant between 1 and R, where R is a large parameter to be chosen

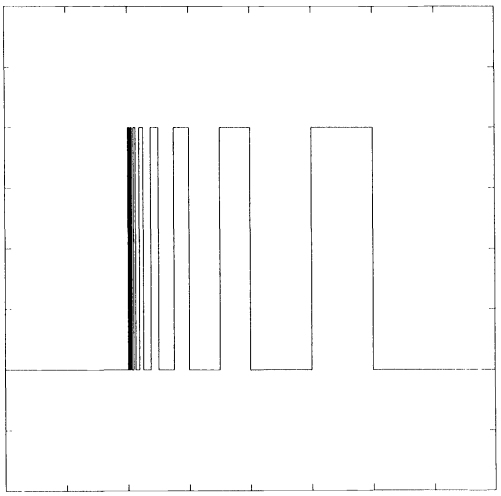


FIG. 2. *A function in* $\mathcal{A}_q^\alpha(L^1([0,1]))$ *for all* $\alpha > 1$ *but not in* $\mathrm{BV}([0,1])$.

later. Between 0 and 1 define $u_0(x)$ by

$$u_0(x) = \sum_{\substack{p/2^k < x \\ k>0,\ 0<p<2^k,\ p \text{ odd}}} \frac{1}{k^r 2^{\beta k}}, \qquad r > 0, \quad \beta > 1.$$

Then it can be shown that u_0 is in any space $\mathcal{A}_q^\alpha(L^1([0,1]))$ containing $\mathcal{A}_{1/r}^\beta(L^1([0,1]))$. Therefore, $u(\cdot,t) \in \mathcal{A}_q^\alpha(L^1([0,R]))$ for the same values of α and q.

Consider the solution $u(\cdot,t)$ of (C) for t between 0 and T for some T when $f(u) = u^2$. The increasing part of u_0 between 0 and 1 spreads out into a series of expansion waves, and there is a shock emanating from the point $(R,0)$ in (x,t) space. For a fixed T, if R is big enough then these waves will not interact. Consider now the solution of

$$\begin{aligned} v_t + g(v)_x &= 0, \qquad x \in \mathbb{R}, \quad t > 0, \\ v(x,0) &= u(x,T), \qquad x \in \mathbb{R}, \end{aligned}$$

with $g(u) = -u^2$. It is easily seen that $v(x,T) = u_0(x)$ for x between 0 and 1, while the rest of $v(x,T)$ consists of constant states and a linear function representing a rarefaction wave. It follows that $v(x,0) = u(x,T)$ cannot have more smoothness than u_0 in the sense of these approximation spaces.

It is interesting to note that $u(\cdot,t)$ is piecewise C^∞ for all positive t and is in the Sobolev space $W^{1,\infty}([0,R])$, yet it is not in the spaces $\mathcal{A}_q^\alpha([0,R])$ if α is large enough.

7. Hamilton-Jacobi equations. A special Hamilton-Jacobi equation in one space dimension is given by

$$\text{(H-J)} \qquad \begin{aligned} w_t + f(w_x) &= 0, \qquad x \in \mathbb{R}, \quad t > 0, \\ w(x,0) &= w_0(x), \qquad x \in \mathbb{R}. \end{aligned}$$

Problems of existence and uniqueness of solutions of (H-J) were solved in papers by M. G. Crandall and P. L. Lions [3], [4], in which they showed that the notion of "viscosity

solution" of (H-J) led to well-posedness. Certain "structural" regularity results are known for solutions of (H-J); see, for example, [14].

The problem (C) can be derived formally from (H-J) by setting $u = w_x$ and differentiating (H-J) with respect to x. This association is more than formal, however, because Crandall and Lions showed that the viscosity solution of (H-J) is the limit as ϵ tends to zero of the solution of (H-J) with the right-hand side replaced by ϵw_{xx} (hence the name "viscosity solution"). The entropy solution of (C) is also the limit as ϵ tends to zero of the equation with the right-hand side replaced with ϵu_{xx} (see, e.g., [15]), so if w_0' is in L^1, then the formal calculations are in fact valid. Thus one can immediately derive the following theorem from the results in §4.

THEOREM 7.1. *Let w_0 have support in $[0,1]$, and assume that there is an $\alpha \in (0,2)$ and a $q \in (0,\infty]$ such that $w_0' \in \mathrm{BV}(\mathbb{R}) \cap \mathcal{A}_q^\alpha(L^1([0,1]))$. Assume also that $f'' \geq 0$, $f(0) = 0$, and that f' and f''' are bounded on Ω (see the comment following Lemma 4.1). Then $w(\cdot,t)$ has support in $I_t = [\inf_{\xi\in\Omega} f'(\xi)t, 1 + \sup_{\xi\in\Omega} f'(\xi)t]$ and $w_x(\cdot,t) \in \mathrm{BV}(\mathbb{R}) \cap \mathcal{A}_q^\alpha(L^1(I_t))$. In particular, when $q = 1/(\alpha+1)$ and $w_0' \in \mathrm{BV}(\mathbb{R}) \cap B_q^\alpha(L^q(I))$, then $w_x(\cdot,t) \in \mathrm{BV}(\mathbb{R}) \cap B_q^\alpha(L^q(I_t))$.*

Acknowledgments. This paper was written after extensive conversations with R. DeVore and V. Popov, and I am deeply indebted to them for their assistance.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] C. DE BOOR, *A Practical Guide to Splines*, Springer, New York, 1978.

[3] M. G. CRANDALL AND P. L. LIONS, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[4] ———, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[5] C. M. DAFERMOS, *Generalized characteristics and the structure of solutions of hyperbolic conservation laws*, Indiana Univ. Math. J., 26 (1977), pp. 1097–1119.

[6] ———, *Regularity and large time behaviour of solutions of a conservation law without convexity*, Proc. Roy. Soc. Edinburgh, 99A (1985), pp. 201–239.

[7] R. A. DEVORE AND V. A. POPOV, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[8] ———, *Interpolation spaces and non-linear approximation*, in Proceedings of the Conference on Interpolation of Operators and Allied Topics in Analysis, Lund, 1986, to appear.

[9] R. J. DIPERNA, *Singularities and oscillations in solutions to conservation laws*, Physica, 12D (1984), pp. 363–368.

[10] ———, *Singularities of solutions of nonlinear hyperbolic systems of conservation laws*, Arch. Rat. Mech. Anal., 60 (1976), pp. 75–100.

[11] H. FEDERER, *Geometric Measure Theory*, Springer, New York, 1969.

[12] J. GLIMM, *Solutions in the large for nonlinear hyperbolic systems of hyperbolic conservation laws*, Comm. Pure Appl. Math., 18 (1965), pp. 697-715.

[13] J. GUCKENHEIMER, *Solving a single conservation law*, in Lecture Notes in Mathematics, 468, Springer-Verlag, Berlin, 1975, pp. 108–134.

[14] R. JENSEN AND P. E. SOUGANIDIS, *A regularity result for viscosity solutions of Hamilton-Jacobi equations in one space dimension*, IMA Preprint Series, 238 (1986).

[15] S. N. KRUŽKOV, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.

[16] T. P. LIU, *Admissible solutions of hyperbolic conservation laws*, Mem. Amer. Math. Soc., 240 (1981).

[17] B. J. LUCIER, *A moving mesh numerical method for hyperbolic conservation laws*, Math. Comp., 46 (1986), pp. 59–69.

[18] O. A. OLEINIK, *Discontinuous solutions of non-linear differential equations*, Usp. Mat. Nauk (N.S.), 12 (1957), pp. 3–73. English translation, Amer. Math. Soc. Transl., Ser. 2, 26, pp. 95–172.

[19] P. PETRUSHEV, *Direct and converse theorems for best spline approximation with free knots and Besov spaces*, C. R. Acad. Bulgare Sci., 39 (1986), pp. 25–28.

[20] ———, *Direct and converse theorems for spline and rational approximation and Besov spaces*, in Proceedings of the Conference on Interpolation of Operators and Allied Topics in Analysis, Lund, 1986, to appear.

[21] D. G. SCHAEFFER, *A regularity theorem for conservation laws*, Adv. in Math., 11 (1973), pp. 368–386.

[22] A. I. VOL'PERT, *The spaces BV and quasilinear equations*, Math. USSR Sbornik, 2 (1967), pp. 225–267.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

002

INVARIANT REGIONS AND GLOBAL ASYMPTOTIC STABILITY IN AN ISOTHERMAL CATALYST*

JOSÉ M. VEGA†

Abstract. A well-known model for the evolution of the (space-dependent) concentration and (lumped) temperature in a porous catalyst is considered. A sequence of invariant regions of the phase space is given, which converges to a globally asymptotically stable region B. Quantitative sufficient conditions are obtained for (the region B to consist of only one point and) the problem to have a (unique) globally asymptotically stable steady state.

Key words. global stability, invariant regions, porous catalysts, isothermal catalysts

AMS(MOS) subject classifications. 35B35, 35B40, 35K57, 80A32

1. Introduction. This paper is concerned with a well-known model (Aris [1]) for the evolution of a single reactant concentration u and of the uniform temperature v in an isothermal catalyst

$$\partial u/\partial t = \Delta u - \phi^2 f(u, v) \quad \text{in } \Omega, \qquad \partial u/\partial n = \sigma(1-u) \qquad \text{on } \partial\Omega, \tag{1.1}$$

$$dv/dt = \lambda\mu(1-v) + \lambda\phi^2 \int_\Omega f(u, v)\, dx. \tag{1.2}$$

Here, Δ is the Laplacian operator and n is the outward unit normal to the boundary of the bounded domain $\Omega \subset \mathbf{R}^p$ ($p = 1, 2,$ or 3). The parameters ϕ^2, σ, λ, and μ are strictly positive.

As it has been frequently pointed out in the literature ([1] and references given therein), the isothermal model (1.1), (1.2) is not unrealistic because temperature is often lumped in practice, due to the high conductivity of the solid catalyst. In fact, such a model is a first approximation, as $\beta \to 0$ and $\nu \to 0$, of the nonisothermal model, in which temperature is spatially distributed, and given by

$$L^{-1}\partial v/\partial t = \Delta v + \beta\phi^2 f(u, v) \quad \text{in } \Omega, \qquad \partial v/\partial n = \nu(1-v) \quad \text{on } \partial\Omega. \tag{1.3}$$

In this limit, the parameters λ and μ of (1.2) are $\lambda = \beta L/V_\Omega$ and $\mu = \nu/S_\Omega\beta$, where V_Ω and S_Ω are the volume and the external area of the domain Ω (see [2]).

The following basic assumptions will be made:

(**H.1**) The domain $\Omega \subset \mathbf{R}^p$ is bounded and (if $p > 1$) it is uniformly of class $C^{2+\alpha}$, for some $0 < \alpha < 1$. Then, it satisfies uniformly *the interior and exterior sphere properties*: there are two constants, $\rho_1 > 0$ and $\rho_2 > 0$, such that, for every point q of $\partial\Omega$, two hyperspheres, S_1 and S_2, of radius ρ_1 and ρ_2, are tangent to $\partial\Omega$ at q and satisfy: $S_1 \subset \Omega$, $\bar{S}_2 \cap \bar{\Omega} = \{q\}$.

(**H.2**) The function $f : [0, \infty[\times [0, \infty[\to \mathbf{R}$ is of class C^1 and there is a continuous function $F : [0, \infty[\to \mathbf{R}$ such that: (i) $f(0, v) = 0$ for all $v \in [0, \infty[$; (ii) $0 < f(u, v) \leqq F(u)$, $|f_u(u, v)| \leqq F(u)$, $0 < f_v(u, v)$ for all $(u, v) \in]0, \infty[\times]0, \infty[$.

* Received by the editors April 7, 1986; accepted for publication (in revised form) July 21, 1987. This research was partially supported by the Spanish Comisión Asesora de Investigación Científica y Técnica, under grant N/r 2291-83.

† Escuela Tecnica Superior de Ingenieros Aeronáuticos, Universidad Politécnica de Madrid, 28040 Madrid, Spain.

Assumption (H.1) is made for some existence and comparison theorems to be applicable. Assumption (H.2) is satisfied by

$$f_1(u, v) = u^m \exp(\gamma - \gamma/v), \quad m \geqq 1, \quad \gamma \geqq 0, \tag{1.4}$$

$$f_2(u, v) = u^m (k+u)^{-r} \exp(\gamma - \gamma/v), \quad m \geqq 1, \quad \gamma \geqq 0, \quad k > 0, \tag{1.5}$$

$$f_3(u, v) = u^m [k \exp(\gamma_a - \gamma_a/v) + u]^{-r} \exp(\gamma - \gamma/v), \tag{1.6}$$

$$m \geqq 1, \quad \gamma \geqq r\gamma_a \geqq 0, \quad k > 0.$$

The Arrhenius reaction rate function f_1 is most frequently used to model thermal effects on the reaction rate (see [1]). The Langmuir-Hinshelwood functions f_2 and f_3 have received a considerable attention in the literature. Function f_2 was first proposed to model carbon monoxide oxidation over platinum catalysts, which is the main reaction in automotive pollution-abatement devices. Further experimental evidence showed that several hydrocarbons, such as ethylene and propylene, follow similar rate laws when oxidized over noble metal catalysts (see [3]).

In this paper, some global asymptotic stability properties of the steady state of (1.1), (1.2) will be obtained. Of course, results in the literature for model (1.1), (1.3) (see [4]–[6]) apply to (1.1), (1.2) after small changes; unfortunately they are rather mild: the steady state of (1.1), (1.3) is globally asymptotically stable (and hence, it is unique) if the parameter ϕ^2 is small enough. A slightly stronger result was proven in [5], but it requires the function f to satisfy $f(u, 0) > 0$ for $u > 0$, and this property does not hold if f is given by (1.4)–(1.6). The results of [4] were obtained by means of a generalized Gronwall inequality. The results of [5], [6] were established by constructing sub- and supersolutions converging to the steady state; the same idea has been used also in the analysis of related reaction-diffusion problems (see, e.g., [7]–[10]).

Our approach is somewhat different, although it is also based on comparison theorems. We shall construct a sequence $\{B_m\}$ of invariant, stable regions of the phase space of (1.1), (1.2), such that every region B_m traps the transient state of the system in a finite time for arbitrary initial conditions. If the sequence $\{B_m\}$ converges to a region of the phase space B in an appropriate uniform sense, then such region is globally asymptotically stable for (1.1), (1.2). Therefore, B contains the nonwandering set of (1.1), (1.2) (i.e., the set of points (u, v) of the phase space of (1.1), (1.2) such that, for every neighborhood of (u, v), $U \subset C(\bar{\Omega}) \times \mathbf{R}$, and every $T > 0$, there are a constant $t > T$ and a point $(u_0, v_0) \in U$ that are such that the solution of (1.1), (1.2), with initial conditions $(u(0), v(0)) = (u_0, v_0)$, satisfies $(u(t), v(t)) \in U$ (see Hirsch [11])). In particular, B contains every (stable or unstable) steady state, periodic, or quasiperiodic solution, $\cdots$, of (1.1), (1.2). If B consists of only one point, then such point is a globally asymptotically stable (and hence, a unique) steady state of (1.1), (1.2). This method of finding globally asymptotically stable invariant regions for nonmonotone flows is similar to that used by Leung [12] in his study of some prey-predator problems; in some sense, the ideas are in the spirit of the work by Keller [13] and Sattinger [14] on semilinear elliptic problems.

In § 2 we shall prove some basic results and state some definitions. In § 3, a sequence of invariant regions of the phase space of (1.1), (1.2), of the type described above, will be obtained. The results of § 3 will be applied in § 4, to obtain some quantitative sufficient conditions for the steady state of (1.1), (1.2) to be globally asymptotically stable for a function f of a rather general type: $f(u, v) = g(u) \exp(\gamma - \gamma/v)$, which includes the particular instances of (1.4) and (1.5). In particular, we shall obtain global asymptotic stability of the steady state if ϕ^2 is

sufficiently small or large, or if the function g is increasing and γ is sufficiently small. As a corollary, some sufficient conditions for the steady state of (1.1), (1.2) to be unique will be obtained. For results on existence and uniqueness of the steady state of (1.1), (1.3), see [4], [6], [15], [16]. It should be pointed out that to prove uniqueness for large ϕ^2 is not an easy task (see [16]).

The following notation will be widely used in the sequel. If Ω is defined as above, and if $u_1, u_2 \in C(\bar{\Omega})$, then $u_1 \leqq u_2$ will mean that $u_1(x) \leqq u_2(x)$ for all $x \in \bar{\Omega}$, and $u_1 < u_2$ will mean that $u_1 \leqq u_2$ and $u_1 \neq u_2$. If $u_1(x) < u_2(x)$ for all $x \in \bar{\Omega}$, then we shall write $u_1 \ll u_2$.

2. Preliminary results and definitions. Let us first consider some basic results concerning the evolution problem (1.1), (1.2), with initial conditions

$$u(x, 0) = \tilde{u}(x) \geqslant 0 \quad \text{for all } x \in \bar{\Omega}, \quad v(0) = \tilde{v} \geqq 0, \tag{2.1}$$

where $\tilde{u} \in C^2(\bar{\Omega})$ and satisfies the boundary condition (1.1). By a (classical) *regular solution* of (1.1), (1.2), (2.1) we shall mean a couple of functions

$$u \in C^{1,0}(\bar{\Omega} \times [0, \infty[) \cap C^{2,1}(\bar{\Omega} \times]0, \infty[), \qquad v \in C^1([0, \infty[),$$

which satisfy (1.1), (1.2), (2.1) pointwise, and are such that $u(\cdot, t) \geqq 0$, $v(t) \geqq 0$ for all $t > 0$. Here, $u \in C^{1,0}$ means that the functions $(x, t) \to u$ and $(x, t) \to Du$ are continuous. $u \in C^{2,1}$ means that $u \in C^{1,0}$ and the functions $(x, t) \to D^2u$ and $(x, t) \to \partial u/\partial t$ are continuous, where Du and D^2u are the matrices of first- and second-order x-derivatives of u. Observe that negative concentrations and temperatures are not allowed since they do not make sense from the physical point of view.

The following consequence of maximum principles will be widely used in the sequel.

LEMMA 2.1. *Let Ω be as in assumption* (H.1), *and let W be a function of $C^{1,0}(\bar{\Omega} \times [0, \infty[) \cap C^{2,1}(\bar{\Omega} \times]0, \infty[)$, such that*

(a) $W(x, 0) \geqq 0$ *for all* $x \in \bar{\Omega}$.

(b) $\partial W/\partial t > \Delta W$ *for all* $(x, t) \in \Omega \times]0, \infty[$ *such that* $W(x, t) < 0$.

(c) $\partial W/\partial n > 0$ *for all* $(x, t) \in \partial\Omega \times]0, \infty[$ *such that* $W(x, t) < 0$.

Then $W(x, t) \geqq 0$ *for all* $(x, t) \in \bar{\Omega} \times [0, \infty[$.

Proof. The result follows by standard arguments, using maximum principles (Protter and Weinberger [17]).

Global existence and uniqueness of solution of (1.1), (1.2), (2.1) will be a consequence of the following a priori bound.

LEMMA 2.2. *Under assumptions* (H.1) *and* (H.2), *let $u = u(x, t)$, $v = v(t)$, be a regular solution of* (1.1), (1.2), (2.1). *Then, there is a constant $\alpha > 0$ and a function $\psi \in C^2(\bar{\Omega})$, such that*

$$u(x, t) \leqq 1 + \psi(x) \exp(-\alpha t) \quad \textit{for all } (x, t) \in \bar{\Omega} \times [0, \infty[.$$

Proof. As is well known, the problem

$$\Delta\psi + \alpha\psi = 0 \quad \text{in } \Omega, \qquad \partial\psi/\partial n + \sigma\psi = 0 \quad \text{on } \partial\Omega, \tag{2.2}$$

has a smallest eigenvalue $\alpha > 0$, and eigenfunctions ψ such that $\psi \gg 0$. Hence, ψ may be chosen to be such that $\psi \geqq 2(\tilde{u} - 1)$, $\psi \gg 0$ and the function $W = W(x, t) = 1 - u(x, t) + [2 \exp(-\alpha t) - \exp(-2\alpha t)]\psi(x)/2$ satisfies $W(\cdot, t) \geqq 0$ for all $t \geqq 0$, as it comes out when Lemma 2.1 is applied.

THEOREM 2.3. *Under assumptions* (H.1) *and* (H.2), *the problem* (1.1), (1.2), (2.1) *has a unique regular solution if* $\tilde{u} \in C^2(\bar{\Omega})$ *and* $\tilde{u}$ *satisfies the boundary condition.*

Proof. For given $\tilde{u}$ and $\tilde{v}$, let ψ be as in Lemma 2.2 and let $k = 1 + \max\{\psi(x): x \in \bar{\Omega}\}$. Then, no regular solution of (1.1), (1.2), (2.1) is affected when f is replaced, in (1.1), (1.2), by another function, $\tilde{f}: \mathbf{R}^2 \to \mathbf{R}$, that is defined by: $\tilde{f}(u, v) = 0$ for $u < 0$, $\tilde{f}(u, v) = f(u, |v|)$ for $0 \leqq u \leqq k$, $\tilde{f}(u, v) = f(k, |v|)$ for $u > k$. Any solution of (1.1), (1.2), (2.1), with f modified as above, is a regular solution of the original problem (the converse is trivially satisfied). That is, if $u \in C^{1,0}(\bar{\Omega} \times [0, \infty[) \cap C^{2,1}(\bar{\Omega},]0, \infty[)$, $v \in C^1([0, \infty[)$ is a solution of the modified problem, then $u(\cdot, t) \geqq 0$ and $v(t) \geqq 0$ for all $t > 0$. Since $\tilde{f}(u, v) \geqq 0$ for all $(u, v) \in \mathbf{R}^2$, (1.2) yields $dv/dt \geqq \lambda\mu(1 - v)$ and $v(t) \geqq 0$ for all $t > 0$; also, $u(\cdot, t) \geqq 0$ for all $t > 0$, as it comes out when Lemma 2.1 is applied to $W = u \exp(-t)$, and it is taken into account that $\tilde{f}(u, v) = 0$ for $u < 0$. Then, we only need to prove the conclusion of the theorem when f is replaced by $\tilde{f}$, and this comes out from standard theory on semilinear equations (e.g., from [18, Cor. 3.3.5] and [19, Lemma 4.2]), when taking into account that $\tilde{f}$ is locally Lipschitz and globally bounded.

The following $\varepsilon - \delta$ stability definitions of the Lyapunov type will be used in the sequel. They are given in terms of the distance d, associated with the norm

$$\|(u, v)\| = \max\{|u(x)|: x \in \bar{\Omega}\} + |v| \quad \text{for } (u, v) \in C(\bar{\Omega}) \times \mathbf{R}.$$

The distance between $(u, v) \in C(\bar{\Omega}) \times \mathbf{R}$ and $B \subset C(\bar{\Omega}) \times \mathbf{R}$ is defined as usually $d[(u, v), B] = \inf\{\|(u - u', v - v')\|: (u', v') \in B\}$. Observe that $C(\bar{\Omega}) \times \mathbf{R}$ includes the phase space of (1.1), (1.2), (2.1).

DEFINITION 2.4. Let $B \subset C(\bar{\Omega}) \times \mathbf{R}$. B is said to be an *invariant region* for the problem (1.1), (1.2), (2.1) if, for any regular solution of the problem, $(u(\cdot, 0), v(0)) \in B$ implies $(u(\cdot, t), v(t)) \in B$ for all $t > 0$. An invariant region B is said to be *stable* if, for every $\varepsilon > 0$, there is a $\delta > 0$ such that for every regular solution of the problem $d[(u(\cdot, 0), v(0)), B] < \delta$ implies $d[(u(\cdot, t), v(t)), B] < \varepsilon$ for all $t > 0$. A region B is said to be *globally asymptotically attracting* if every regular solution of the problem satisfies $d[(u(\cdot, t), v(t)), B] \to 0$ as $t \to \infty$. An invariant region B is said to be *globally asymptotically stable* if it is stable and globally asymptotically attracting. A region B is said to be *globally finitely attracting* if, for every regular solution of the problem, there is a constant $T < \infty$ such that $(u(\cdot, t), v(t)) \in B$ for all $t \geqq T$.

The concept of globally finitely attracting region and the following lemma will be used in § 3.

LEMMA 2.5. *Let the sequence of regions* $\{B_m\}$ *and the region* $B = \cap\{B_m: m \in \mathbf{N}\}$, *of* $C(\bar{\Omega}) \times \mathbf{R}$, *be such that*

(a) *Every* B_m *is invariant, and globally finitely attracting for* (1.1), (1.2), (2.1);

(b) *For every* $m \in \mathbf{N}$, *there are two constants,* $\varepsilon_m > 0$ *and* $\delta_m > 0$, *such that* $N(B, \delta_m) \subset B_m \subset N(B, \varepsilon_m)$, *where*

$$N(B, \delta) = \{(u, v) \in C(\bar{\Omega}) \times \mathbf{R}: d[(u, v), B] < \delta\};$$

(c) $\varepsilon_m \to 0$ *as* $m \to \infty$.

Then the region B is invariant and globally asymptotically stable for the problem.

Proof. Since $B = \cap\{B_m: m \in \mathbf{N}\}$, the region B is clearly invariant. B is stable since for every $\varepsilon > 0$ there is an $m \in \mathbf{N}$ such that $\varepsilon_m < \varepsilon$; then the definition of stable region is satisfied with $\delta = \delta_m$. Finally, B is globally asymptotically attracting since for every $\varepsilon > 0$ there is a constant T such that $d[(u(\cdot, t), v(t)), B] < \varepsilon$ for all $t \geqq T$. To see that, take m such that $\varepsilon_m < \varepsilon$ and take into account that B_m is globally finitely attracting and $B_m \subset N(B, \varepsilon_m) \subset N(B, \varepsilon)$.

3. Invariant regions. In this section, we obtain a sequence of regions satisfying the hypothesis of Lemma 2.5, which leads to a globally asymptotically stable region of the phase space of (1.1), (1.2), (2.1).

Let $\{\alpha^m\}$ be a strictly decreasing sequence of real numbers, such that $\alpha^m \to 1$ as $m \to \infty$ and $\alpha^0(\alpha^1 - 1) \leqq M^{-1}\alpha^1 f(\alpha^0, 1/\alpha^0)$, where the constant $M > 0$ is defined below. Let the sequence $\{\alpha_m\}$ be defined by $\alpha_m = 1/\alpha^m$ for all $m \in \mathbf{N}$. From assumption (H.2) (see Introduction), it turns out that there is a constant $M > 0$ and a function $h : [0, \alpha^0] \times [\alpha_0, \infty[\to \mathbf{R}$, of class C^1 and bounded, such that $h \geqq 0$, $\partial h/\partial u \geqq 0$, $\partial h/\partial v \geqq 0$, $-\partial h/\partial u \leqq \partial f/\partial u < M$ for all $(u, v) \in [0, \alpha^0] \times [\alpha_0, \infty[$.

We consider the sequence of regions $\{B_m\} \subset C(\bar{\Omega}) \times \mathbf{R}$, defined by

$$(3.1)\quad B_m = \left\{(u, v) \in C(\bar{\Omega}) \times \mathbf{R}:\ u_m \leqq u \leqq u^m,\ G_m \leqq v + \lambda \int_\Omega u\, dx \leqq G^m,\ v_m \leqq v \leqq v^m\right\},$$

where u_0, u^0, G_0, G^0, v_0, and v^0 are

$$(3.2)\quad u_0 = 0, \quad G_0 = \alpha_0 - \mu^{-1}\sigma S_\Omega(\alpha^0 - 1), \quad v_0 = \alpha_0, \quad u^0 = \alpha^0,$$

$$(3.3)\quad \begin{aligned} G^0 &= \alpha^0(1 + \lambda V_\Omega) + \mu^{-1}\sigma S_\Omega, \\ v^0 &= \alpha^0 + \mu^{-1}\phi^2 V_\Omega \sup\{f(\alpha^0, v) + h(\alpha^0, v):\ v \geqq \alpha_0\} \end{aligned}$$

(V_Ω and S_Ω are the volume of Ω and the area of $\partial\Omega$, respectively), and where u_m, u^m, G_m, G^m, v_m, and v^m ($m \geqq 1$) are defined, inductively, by

$$(3.4)\quad \begin{aligned} \Delta u_m - \phi^2 M u_m &= \alpha_m \phi^2 [f(u_{m-1}, v^{m-1}) - M u_{m-1}] \quad \text{in } \Omega, \\ \partial u_m/\partial n &= \sigma(\alpha_m - u_m) \quad \text{on } \partial\Omega, \end{aligned}$$

$$(3.5)\quad \begin{aligned} \Delta u^m - \phi^2 M u^m &= \alpha^m \phi^2 [f(u^{m-1}, v_{m-1}) - M u^{m-1}] \quad \text{in } \Omega, \\ \partial u^m/\partial n &= \sigma(\alpha^m - u^m) \quad \text{on } \partial\Omega, \end{aligned}$$

$$(3.6)\quad \begin{aligned} G_m &= \alpha_m + \lambda \int_\Omega u_m\, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1 - u^m)\, ds, \\ G^m &= \alpha^m + \lambda \int_\Omega u^m\, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1 - u_m)\, ds, \end{aligned}$$

$$(3.7)\quad v_m = \max\left\{\alpha_m, G_m - \lambda \int_\Omega u^m\, dx, w_m\right\}, \qquad v^m = \min\left\{G^m - \lambda \int_\Omega u_m\, dx, w^m\right\}$$

with

$$(3.8)\quad w_m = \alpha_m + \mu^{-1}\phi^2 \int_\Omega [f(u_m, v_{m-1}) + h(u_m, v_{m-1}) - h(u^m, v^{m-1})]\, dx,$$

$$(3.9)\quad w^m = \alpha^m + \mu^{-1}\phi^2 \int_\Omega [f(u^m, v^{m-1}) + h(u^m, v^{m-1}) - h(u_m, v_{m-1})]\, dx.$$

LEMMA 3.1. *Let $m \geqq 0$ be an integer. If a regular solution of* (1.1), (1.2), (2.1) *satisfies, for all $t \geqq 0$,*

$$(3.10)\quad u_0 \leqq u(\cdot, t) \leqq u^0, \qquad \alpha_0 \leqq v(t) \quad \text{if } m = 0, \quad \text{or}$$

$$(3.11)\quad u_m \leqq u(\cdot, t) \leqq u^m, \qquad v_{m-1} \leqq v(t) \leqq v^{m-1} \quad \text{if } m \geqq 1,$$

then there is a constant T such that for all $t \geqq T$,

$$G_m \leqq v(t) + \lambda \int_\Omega u(x, t)\, dx \leqq G^m, \qquad v_m \leqq v(t) \leqq v^m. \tag{3.12}$$

If, in addition, the inequalities (3.12) *hold for $t = 0$, then they also hold for all $t > 0$.*

Proof. By using (3.10) or (3.11), the time derivative of $G(t) = v(t) + \lambda \int_\Omega u(x, t)\, dx$,

$$dG/dt = \lambda\mu(1 - G) + \lambda^2\mu \int_\Omega u\, dx + \lambda\sigma \int_{\partial\Omega} (1 - u)\, ds,$$

and dv/dt are easily seen to satisfy, for all $t \geqq 0$,

$$\lambda\mu(G_m - G + 1 - \alpha_m) \leqq dG/dt \leqq \lambda\mu(G^m - G + 1 - \alpha^m) \quad \text{for } m \geqq 0,$$

$$\lambda\mu(1 - v) \leqq dv/dt \leqq \lambda\mu(v^0 - v + 1 - \alpha^0),$$

$$\lambda\mu(w_m - v + 1 - \alpha_m) \leqq dv/dt \leqq \lambda\mu(w^m - v + 1 - \alpha^m) \quad \text{for } m \geqq 1.$$

From these inequalities, the conclusion of the lemma readily follows.

LEMMA 3.2. *The sequences defined by* (3.1)-(3.7) *satisfy, for all $m \in \mathbf{N}$:*

A. $u_m \ll u_{m+1} \ll u^{m+1} \ll u^m$, $G_m < G_{m+1} < G^{m+1} < G^m$, $v_m < v_{m+1} < v^{m+1} < v^m$.

B. *B_m is an invariant region for the problem* (1.1), (1.2), (2.1).

C. *B_m is a globally, finitely attracting region for* (1.1), (1.2), (2.1).

Proof. An induction argument will be used in the three cases. It will be proved that the required property holds for $m = 0$ and that it is satisfied for $m = p$ if it holds for $m = p - 1$.

A. Both steps of the induction argument are easily accomplished by means of maximum principles.

B. To prove that B_0 is invariant, observe that if a regular solution of (1.1), (1.2), (2.1) is such that $(u(\cdot, 0), v(0)) \in B_0$, then it satisfies, for all $t \geqq 0$: (i) $u(\cdot, t) \geqq u_0 = 0$ (definition of regular solution); (ii) $v(t) \geqq v_0 = \alpha_0$ (use the inequality $dv/dt \geqq \lambda\mu(1 - v)$); (iii) $u(\cdot, t) \leqq u^0 = \alpha^0$ (apply Lemma 2.1 with $W = \alpha^0 - u$); and (iv) $G_0 \leqq v(t) + \lambda \int_\Omega u(x, t)\, dx \leqq G^0$, $v(t) \leqq v^0$ (Lemma 3.1). In the same way, if B_{p-1} is invariant and if $(u(\cdot, 0), v(0)) \in B_p \subset B_{p-1}$, then for all $t \geqq 0$, $(u(\cdot, t), v(t)) \subset B_{p-1}$ and (i) $u_p \leqq u(\cdot, t) \leqq u^p$ (apply Lemma 2.1 with $W = u - u_p$ and with $W = u^p - u$), and (ii) $G_p \leqq v(t) + \lambda \int_\Omega u(x, t)\, dx \leqq G^p$, $v_p \leqq v(t) \leqq v^p$ (Lemma 3.1).

C. To prove that B_0 is globally finitely attracting, observe that any regular solution of (1.1), (1.2), (2.1) satisfies, for some finite constants, T_1, T_2, and T_3 (i) $0 = u_0 \leqq u(\cdot, t) \leqq u^0 = \alpha^0$ for all $t \geqq T_1$ (Lemma 2.2); (ii) $v(t) \geqq v_0 = \alpha_0$ for all $t \geqq T_2$ (use the inequality $dv/dt \geqq \lambda\mu(1 - v)$); and (iii) $G_0 \leqq v(t) + \lambda \int_\Omega u(x, t)\, dx \leqq G^0$, $v_0 \leqq v(t) \leqq v^0$ for all $t \geqq T_3$ (take the time variable $t = t - \max\{T_1, T_2\}$ and apply Lemma 3.1).

Now, we assume that B_{p-1} satisfies property C and prove that B_p also satisfies it. Let (u, v) be a regular solution of (1.1), (1.2), (2.1). By taking an appropriate origin of the time scale, we may assume that $u_{p-1} \leqq u(\cdot, t) \leqq u^{p-1}$ and $v_{p-1} \leqq v(t) \leqq v^{p-1}$ for all $t \geqq 0$. Then there are finite constants, T_1 and T_2, such that (i) $u_p \leqq u(\cdot, t) \leqq u^p$ for all $t \geqq T_1$ (apply Lemma 2.1 with $W = u + \psi_p \exp(-\alpha t) - \alpha^p u_p$ and with $W = \alpha_p u^p + \psi^p \exp(-\alpha t) - u$, where $\alpha > 0$ is the smallest eigenvalue of (2.2), and $\psi_p \geqq 0$ and $\psi^p \geqq 0$ are eigenfunctions such that $\psi_p \geqq \alpha^p u_p - u(\cdot, 0)$ and $\psi^p \geqq u(\cdot, 0) - \alpha_p u^p$); (ii) $G_p \leqq v(t) + \lambda \int_\Omega u(x, t)\, dx \leqq G^p$, $v_p \leqq v(t) \leqq v^p$ for all $t \geqq T_2$ (take the time variable $t = t - T_1$ and apply Lemma 3.1).

THEOREM 3.3. A. *The sequences defined by* (3.2)-(3.7) *satisfy* $u_m \to u_*$, $u^m \to u^*$, *uniformly in* $\bar{\Omega}$; $G_m \to G_*$, $G^m \to G^*$, $v_m \to v_*$, $v^m \to v^*$, *as* $m \to \infty$, *where* $u_*, u^* \in C^2(\bar{\Omega})$, G_*, G^*, v_*, *and* v^* *satisfy*

$$\Delta u_* = \phi^2 f(u_*, v^*) \quad \text{in } \Omega, \qquad \partial u_*/\partial n = \sigma(1-u_*) \quad \text{on } \partial\Omega, \tag{3.13}$$

$$\Delta u^* = \phi^2 f(u^*, v_*) \quad \text{in } \Omega, \qquad \partial u^*/\partial n = \sigma(1-u^*) \quad \text{on } \partial\Omega, \tag{3.14}$$

$$\begin{aligned} G_* &= 1+\lambda \int_\Omega u_*\, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1-u^*)\, ds, \\ G^* &= 1+\lambda \int_\Omega u^*\, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1-u_*)\, ds, \end{aligned} \tag{3.15}$$

$$v_* = \max\left\{1,\ G_* - \lambda \int_\Omega u^*\, dx,\ 1+\mu^{-1}\phi^2 \int_\Omega [f(u_*, v_*)+h(u_*, v_*)-h(u^*, v^*)]\, dx\right\}, \tag{3.16}$$

$$v^* = \min\left\{G^* - \lambda \int_\Omega u_*\, dx,\ 1+\mu^{-1}\phi^2 \int_\Omega [f(u^*, v^*)+h(u^*, v^*)-h(u_*, v_*)]\, dx\right\}, \tag{3.17}$$

$$0 \ll u_* \leqq u^* \ll 1, \quad 1 \leqq G_* \leqq G^* < \infty, \quad 1 \leqq v_* \leqq v^* < \infty. \tag{3.18}$$

B. *The region*

$$B = \left\{(u, v) \in C(\bar{\Omega}) \times \mathbf{R}:\ u_* \leqq u \leqq u^*,\ G_* \leqq v + \lambda \int_\Omega u\, dx \leqq G^*,\ v_* \leqq v \leqq v^*\right\}$$

is invariant, and globally asymptotically stable for the problem (1.1), (1.2), (2.1).

Proof. A. The monotone, bounded sequences $\{G_m\}$, $\{G^m\}$, $\{v_m\}$, and $\{v^m\}$ are convergent, and their limits satisfy (3.18) (Lemma 3.2A). In the same way, the monotone, bounded sequences $\{u_m\}$ and $\{u^m\}$ are pointwise-convergent to some functions u_* and u^* satisfying $0 \ll u_* \leqq u^*$. By means of elliptic estimates, it may be seen that u_* and u^* are twice continuously differentiable and satisfy (3.13), (3.14), and that the convergence is uniform in $\bar{\Omega}$ (only slight modifications are necessary in the proof of Theorem 2.1 of [14], or in the proof of Theorem 10.3 of [20]). Then (3.15)-(3.17) are obtained as limits of (3.6), (3.7). The inequality $u^* \ll 1$ is easily obtained when maximum principles are applied to (3.14).

B. The sequence $\{B_m\}$ satisfies the hypothesis (a) of Lemma 2.5 (Lemma 3.2). Hypothesis (c) is also satisfied if δ_m and ε_m are

$$\begin{aligned} (1+\lambda V_\Omega)\delta_m = \min\{&\min\{u_* - u_m:\ x \in \bar{\Omega}\},\ \min\{u^m - u^*:\ x \in \bar{\Omega}\}, \\ &G_* - G_m,\ G^m - G^*,\ v_* - v_m,\ v^m - v^*\}, \\ \varepsilon_m = 3(1+\lambda V_\Omega)\max\{&\max\{u_* - u_m:\ x \in \bar{\Omega}\},\ \max\{u^m - u^*:\ x \in \bar{\Omega}\}, \\ &G_* - G_m,\ G^m - G^*,\ v_* - v_m,\ v^m - v^*\}. \end{aligned}$$

Observe that $\delta_m > 0$ for $m = 0, 1, \cdots$, as it comes out from the inequalities $u_m \ll u_* \leqq u^* \ll u^m$, $G_m < G_* \leqq G^* < G^m$, $v_m < v_* \leqq v^* < v^m$ (for $m = 0, 1, \cdots$), which are easily obtained from Lemma 3.2A.

Then, we only need to prove that hypothesis (b) is also satisfied. To this end, observe that.

(i) If $(u, v) \in N(B, \delta_m)$, then there is $(u', v') \in B$ such that

$$d[(u, v), (u', v')] = \max \{|u - u'|: x \in \bar{\Omega}\} + |v - v'| < \delta_m.$$

Hence, $u_* \leqq u' \leqq u^*$, $G_* \leqq v' + \lambda \int_\Omega u'\, dx \leqq G^*$, $v_* \leqq v' \leqq v^*$ and

$$u^m - u \geqq (u^m - u^*) + (u^* - u') - |u - u'| \geqq (1 + \lambda V_\Omega)\delta_m + 0 - \delta_m \geqq 0,$$

$$G^m - v - \lambda \int_\Omega u\, dx \geqq (G^m - G^*) + \left(G^* - v' - \lambda \int_\Omega u'\, dx\right) - \left(|v - v'| + \lambda \int_\Omega |u - u'|\, dx\right)$$

$$\geqq (1 + \lambda V_\Omega)\delta_m + 0 - (1 + \lambda V_\Omega)\delta_m \geqq 0,$$

$$v^m - v \geqq (v^m - v^*) + (v^* - v') - |v - v'| \geqq (1 + \lambda V_\Omega)\delta_m + 0 - \delta_m \geqq 0.$$

Similarly, it is easily seen that $u - u_m \geqq 0$, $v + \lambda \int_\Omega u\, dx - G_m \geqq 0$ and $v - v_m \geqq 0$. Hence, $(u, v) \in B_m$.

(ii) If $(u, v) \in B_m$ and if $(u', v') \in C(\bar{\Omega}) \times \mathbf{R}$ is given by

$$u'(x) = \max \{u_*(x), \min \{u(x), u^*(x)\}\} \quad \text{for } x \in \bar{\Omega},$$

$$v' = \max \left\{ v_*, G_* - \lambda \int_\Omega u'\, dx, \min \left\{ v, v^*, G^* - \lambda \int_\Omega u'\, dx \right\} \right\},$$

then $(u', v') \in B$ and $d[(u, v), (u', v')] \leqq 2\varepsilon_m/3 < \varepsilon_m$, as is easily seen. Therefore, $(u, v) \in N(B, \varepsilon_m)$.

Remark 3.4. Some remarks about the results above are in order.

A. It is easily seen, by means of an induction argument, that for every solution of (3.13)-(3.18), $(u_*, u^*, G_*, G^*, v_*, v^*)$, the sequence defined by (3.2)-(3.7) satisfies

$$u_m \ll u_* \leqq u^* \ll u^m, \qquad G_m < G_* \leqq G^* < G^m, \qquad v_m < v_* \leqq v^* < v^m, \tag{3.19}$$

for all $m \in \mathbf{N}$. Therefore, the solution of (3.13)-(3.18) that is approached as $m \to \infty$ by the sequence (3.2)-(3.7), $(\tilde{u}_*, \tilde{u}^*, \tilde{G}_*, \tilde{G}^*, \tilde{v}_*, \tilde{v}^*)$, is maximal in the following sense: any other solution of (3.13)-(3.18) is such that $\tilde{u}_* \leqq u_* \leqq u^* \leqq \tilde{u}^*$, $\tilde{G}_* \leqq G_* \leqq G^* \leqq \tilde{G}^*$, $\tilde{v}_* \leqq v_* \leqq v^* \leqq \tilde{v}^*$. Since such maximal solution of (3.13)-(3.18) is necessarily unique, the region B of Theorem 3.3 is independent of the choice of the sequence $\{\alpha^m\}$ and of the constant M. Furthermore, if one takes $\alpha_m = \alpha^m = 1$ for all $m \in \mathbf{N}$ in (3.1)-(3.9), the following sequence of regions is obtained

$$B_m = \left\{ (u, v) \in C(\bar{\Omega}) \times \mathbf{R}: u_m \leqq u \leqq u^m, G_m \leqq v + \lambda \int_\Omega u\, dx \leqq G^m, v_m \leqq v \leqq v^m \right\}, \tag{3.20}$$

where u_0, u^0, G_0, G^0, v_0, and v^0 are

$$u_0 = 0, \qquad G_0 = v_0 = 1, \qquad u^0 = 1, \qquad G^0 = 1 + \lambda V_\Omega + \mu^{-1}\sigma S_\Omega, \tag{3.21}$$

$$v^0 = 1 + \mu^{-1}\phi^2 V_\Omega \sup \{f(1, v) + h(1, v): v \geqq 1\}, \tag{3.22}$$

and where u_m, u^m, G_m, G^m, v_m, and v^m $(m \geqq 1)$ are defined inductively by

$$\begin{aligned} \Delta u_m - \phi^2 M u_m &= \phi^2[f(u_{m-1}, v^{m-1}) - M u_{m-1}] \quad \text{in } \Omega, \\ \partial u_m/\partial n &= \sigma(1-u_m) \quad \text{on } \partial\Omega, \end{aligned} \tag{3.23}$$

$$\begin{aligned} \Delta u^m - \phi^2 M u^m &= \phi^2[f(u^{m-1}, v_{m-1}) - M u^{m-1}] \quad \text{in } \Omega, \\ \partial u^m/\partial n &= \sigma(1-u^m) \quad \text{on } \partial\Omega, \end{aligned} \tag{3.24}$$

$$\begin{aligned} G_m &= 1 + \lambda \int_\Omega u_m \, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1-u^m) \, ds, \\ G^m &= 1 + \int_\Omega u^m \, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1-u_m) \, ds, \end{aligned} \tag{3.25}$$

$$\begin{aligned} v_m = \max \Big\{ & 1, \; G_m - \lambda \int_\Omega u^m \, dx, \\ & 1 + \mu^{-1}\phi^2 \int_\Omega [f(u_m, v_{m-1}) + h(u_m, v_{m-1}) - h(u^m, v^{m-1})] \, dx \Big\}, \end{aligned} \tag{3.26}$$

$$\begin{aligned} v^m = \min \Big\{ & G^m - \lambda \int_\Omega u_m \, dx, \\ & 1 + \mu^{-1}\phi^2 \int_\Omega [f(u^m, v^{m-1}) + h(u^m, v^{m-1}) - h(u_m, v_{m-1})] \, dx \Big\}. \end{aligned} \tag{3.27}$$

The sequence defined by (3.21)-(3.27) is such that (i) it approaches a solution of (3.13)-(3.18) as $m \to \infty$ (since it is seen to satisfy Lemma 3.2A and Theorem 3.3A), and (ii) it satisfies (3.19) for all $m \in \mathbf{N}$ and for every solution of (3.13)-(3.18) (to prove it, use an induction argument, as above). Hence such sequence also approaches the maximal solution of (3.13)-(3.18) as $m \to \infty$, and the region B of Theorem 3.3 may be obtained as the limit of the sequence of regions defined by (3.20), which may be easily computed (numerically in general) from the linear problems (3.23)-(3.27).

B. As it was mentioned in § 1, since the region B of Theorem 3.3 is globally asymptotically stable, it contains the nonwandering set of (1.1), (1.2), (2.1), and the same is true for any of the regions B_m defined by (3.20)-(3.27) (since $B \subset B_m$ for all $m \in \mathbf{N}$, as it was seen in remark A above). In particular every (stable or unstable) steady state of (1.1), (1.2) is included in B.

C. If every solution of (3.13)-(3.18) satisfies $u_* = u^*$ and $v_* = v^*$, then the region B of Theorem 3.3 is a singleton, $B = \{(u_s, v_s)\}$, and (u_s, v_s) is a globally asymptotically stable steady state of (1.1), (1.2); in addition, (u_s, v_s) is the unique steady state of (1.1), (1.2), as it comes out from remark B above. Observe also that (3.13)-(3.18) has a unique solution in this case. This result will be used in the next section to obtain quantitative, sufficient conditions for global asymptotic stability and uniqueness of the steady state of (1.1), (1.2).

4. Global asymptotic stability of the steady state. In this section, we shall obtain sufficient conditions for global asymptotic stability of the steady state of (1.1), (1.2), (2.1), when the function f is given by

$$f(u, v) = g(u) \exp(\gamma - \gamma/v), \tag{4.1}$$

where $g : [0, \infty[\to \mathbf{R}$ is a C^1-function satisfying

$$g(0) = 0, \qquad g(u) > 0 \quad \text{for all } u > 0. \tag{4.2}$$

Particular instances of such form of f are those in (1.4), (1.5). Some additional assumptions about the function g will be considered below, when needed.

In order to avoid too many involved expressions, we shall obtain only reasonably good sufficient conditions for global stability (and not the best ones that can be obtained from the results of § 3).

The role of the parameters ϕ^2, λ, and σ deserves some attention. The Damköhler number ϕ^2 is the basic parameter; the steady-state solutions of (1.1), (1.2), for example, are usually represented by the curve $\eta - \phi^2$, where η is a significant functional of the steady state, i.e.,

$$\eta = \int_\Omega f(u_s(x), v_s)\, dx / V_\Omega f(1, 1),$$

which is called the effectiveness factor (see [1]). Below, we shall prove that the steady state is globally asymptotically stable (i) if ϕ^2 is sufficiently small or large, for fixed values of the remaining parameters; and (ii) for arbitrary values of ϕ^2 if the parameter γ is sufficiently small and the function g is increasing. The parameter λ is a Lewis number; increasing values of λ are expected to make any steady state of (1.1), (1.2) more and more linearly unstable (i.e., to increase the growth rate of the linear stability analysis). This has been shown to be true for lumped chemically reacting systems (see [21]), and for some distributed systems (such as (1.1), (1.2) if $f(u, v) = u \exp(\gamma - \gamma/v)$; see [2]). Observe that the steady-state solutions of (1.1), (1.2) do not depend on λ. Some of the results below will be independent of λ (they will be valid for $0 < \lambda < \infty$), and some others (depending on λ) will be quite useful for small values of λ. The Sherwood number σ is usually fairly large (see [1]). Some emphasis will be put on obtaining results that are significant as $\sigma \to \infty$ (see, e.g., Theorems 4.4 and 4.5).

Let us assume that the domain Ω satisfies assumption (H.1) (see Introduction). If the function f is as defined by (4.1), then Theorem 3.3 applies. The system (3.13)–(3.18) may be written as

$$\Delta u_* = \phi^2 g(u_*) \exp(\gamma - \gamma/v^*) \quad \text{in } \Omega, \qquad \partial u_*/\partial n = \sigma(1 - u_*) \quad \text{on } \partial\Omega, \tag{4.3}$$

$$\Delta u^* = \phi^2 g(u^*) \exp(\gamma - \gamma/v_*) \quad \text{in } \Omega, \qquad \partial u^*/\partial n = \sigma(1 - u^*) \quad \text{on } \partial\Omega, \tag{4.4}$$

$$\begin{aligned} v_* = 1 + \max\Big\{ 0, -\lambda \int_\Omega (u^* - u_*)\, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1 - u^*)\, ds, \\ \mu^{-1}\phi^2 \Big[\exp(\gamma - \gamma/v_*) \int_\Omega (g(u_*) + h(u_*))\, dx \\ - \exp(\gamma - \gamma/v^*) \int_\Omega h(u^*)\, dx \Big] \Big\}, \end{aligned} \tag{4.5}$$

$$\begin{aligned} v^* = 1 + \min\Big\{ \lambda \int_\Omega (u^* - u_*)\, dx + \mu^{-1}\sigma \int_{\partial\Omega} (1 - u_*)\, ds, \\ \mu^{-1}\phi^2 \Big[\exp(\gamma - \gamma/v^*) \int_\Omega (g(u^*) + h(u^*))\, dx \\ - \exp(\gamma - \gamma/v_*) \int_\Omega h(u_*)\, dx \Big] \Big\}, \end{aligned} \tag{4.6}$$

$$0 \ll u_* \leqq u^* \ll 1, \qquad 1 \leqq v_* \leqq v^* < \infty, \tag{4.7}$$

where $h = [0, 1] \to \mathbf{R}$ is a C^1-function satisfying

$$h'(u) \geqq 0, \; g'(u) + h'(u) \geqq 0 \quad \text{for all } 0 \leqq u \leqq 1. \tag{4.8}$$

The function h may be chosen to be such that

$$k_1 = \max\{0, \max\{-g'(u): 0 \leqq u \leqq 1\}\} = \max\{h'(u): 0 \leqq u \leqq 1\}. \tag{4.9}$$

The main idea to be used in the sequel is the following. According to Remark 3.4C, if every solution of (4.3)-(4.7) satisfies

$$u_* = u^*, \qquad v_* = v^*, \tag{4.10}$$

then (1.1), (1.2), (2.1) possess a unique steady state, which is globally asymptotically stable.

THEOREM 4.1 (*Global asymptotic stability for small* ϕ^2). *Under the assumptions above,* (1.1), (1.2), (2.1) *has a unique steady state, which is globally asymptotically stable if* ϕ^2 *satisfies*

$$\phi^2 \exp\gamma < \alpha/k_1, \tag{4.11}$$

and one of the following inequalities:

$$\gamma V_\Omega\{k_2(2k_1+k_5)[k_3+k_1k_4\phi^2\exp\gamma/(\alpha-k_1\phi^2\exp\gamma)]\phi^2 \cdot \exp\gamma + k_2 + 2k_6\}\phi^2\exp\gamma \leqq \mu, \tag{4.12}$$

$$\gamma k_2(2\lambda\mu V_\Omega + \sigma S_\Omega)[k_3+k_1k_4\phi^2\exp\gamma/(\alpha-k_1\phi^2\exp\gamma)]\phi^2\exp\gamma \leqq \mu, \tag{4.13}$$

$$\gamma k_2 V_\Omega\{1+(2\lambda\mu+k_5\phi^2\exp\gamma)[k_3+k_1k_4\phi^2\exp\gamma/(\alpha-k_1\phi^2\exp\gamma)]\phi^2 \cdot \exp\gamma \leqq \mu, \tag{4.14}$$

where α, k_1, k_2, k_3, *and* k_4 *are as in Lemma* A.1 (*see Appendix*), *and*

$$k_5 = \max\{g'(u): 0 \leqq u \leqq 1\}, \qquad k_6 = \max\{h(u): 0 \leqq u \leqq 1\}.$$

Proof. We shall prove that if (4.11) and one of the inequalities (4.12)-(4.14) hold, then every solution of (4.3)-(4.7) satisfies (4.10). To this end, observe that if (4.11) holds, then u_* and u^* satisfy (Lemma A.1)

$$u^* - u_* \leqq k_2[k_3+k_1k_4\phi^2\exp\gamma/(\alpha-k_1\phi^2\exp\gamma)]\phi^2\exp\gamma[1-\exp(-\gamma\xi)], \tag{4.15}$$

where

$$\xi = 1/v_* - 1/v^* \tag{4.16}$$

is such that

$$0 \leqq \xi \leqq 1, \quad \xi/(1-\xi) \leqq v_*^2\xi/(1-v_*\xi) = v^* - v_*, \tag{4.17}$$

as it comes out from (4.7). Subtraction of (4.5) from (4.6) yields

$$v^* - v_* \leqq \mu^{-1}\phi^2\Big[\exp(\gamma-\gamma/v^*)\int_\Omega (g(u^*)+2h(u^*))\,dx - \exp(\gamma-\gamma/v_*)\int_\Omega (g(u_*)+2h(u_*))\,dx\Big], \tag{4.18}$$

$$v^* - v_* \leqq 2\lambda\int_\Omega (u^*-u_*)\,dx + \mu^{-1}\sigma\int_{\partial\Omega}(u^*-u_*)\,ds. \tag{4.19}$$

Integration over Ω in (4.3) and (4.4) and application of Green's identity yield

$$\sigma\int_{\partial\Omega}(1-u_*)\,ds = \phi^2\exp(\gamma-\gamma/v^*)\int_\Omega g(u_*)\,dx, \tag{4.20}$$

$$\sigma\int_{\partial\Omega}(1-u^*)\,ds = \phi^2\exp(\gamma-\gamma/v_*)\int_\Omega g(u^*)\,dx. \tag{4.21}$$

Substraction of (4.21) from (4.20) and substitution in (4.19) lead to

$$v^* - v_* \leqq 2\lambda \int_\Omega (u^* - u_*) + \mu^{-1}\phi^2 \Big[\exp(\gamma - \gamma/v^*) \int_\Omega g(u_*)\, dx - \exp(\gamma - \gamma/v_*) \int_\Omega g(u^*)\, dx \Big]. \tag{4.22}$$

Finally, after substitution of (4.15)-(4.17) in (4.18), (4.19), and (4.22), the following inequalities are obtained:

$$\xi/(1-\xi) \leqq A_i[1 - \exp(-\gamma\xi)] \quad \text{for } i = 1, 2, \text{ and } 3, \tag{4.23}$$

where $\gamma\mu A_1$, $\gamma\mu A_2$, and $\gamma\mu A_3$ are the first members of (4.12), (4.13), and (4.14). If one of the inequalities (4.12)-(4.14) is satisfied, then $\xi = 0$ (i.e., $v_* = v^*$), as it comes out from (4.23), $u_* = u^*$ (Lemma A.1) and the conclusion of the theorems follows.

Remark. Condition (4.12) does not depend on λ, and it is more stringent than (4.14) if $k_1 \neq 0$ and λ is sufficiently small. If σ is sufficiently small, condition (4.14) is more stringent than (4.13).

THEOREM 4.2 (*Global asymptotic stability for all* $\phi^2 > 0$). *If, in addition to the assumptions of Theorem* 4.1, *the function g satisfies condition* (A.6) *of Lemma* A.2 (*see Appendix*), *then, for all* $\phi^2 > 0$, (1.1), (1.2), (2.1) *have a unique steady state, which is globally asymptotically stable, provided that* γ *satisfies one of the following inequalities* (*see Fig.* 1):

$$k_7\gamma\sigma S_\Omega/\mu \leqq 1/(1 + 2\lambda\mu V_\Omega/\sigma S_\Omega), \tag{4.24}$$

$$\gamma\sigma S_\Omega/\mu \leqq 1/(1 + k_7\lambda\mu V_\Omega/\sigma S_\Omega), \tag{4.25}$$

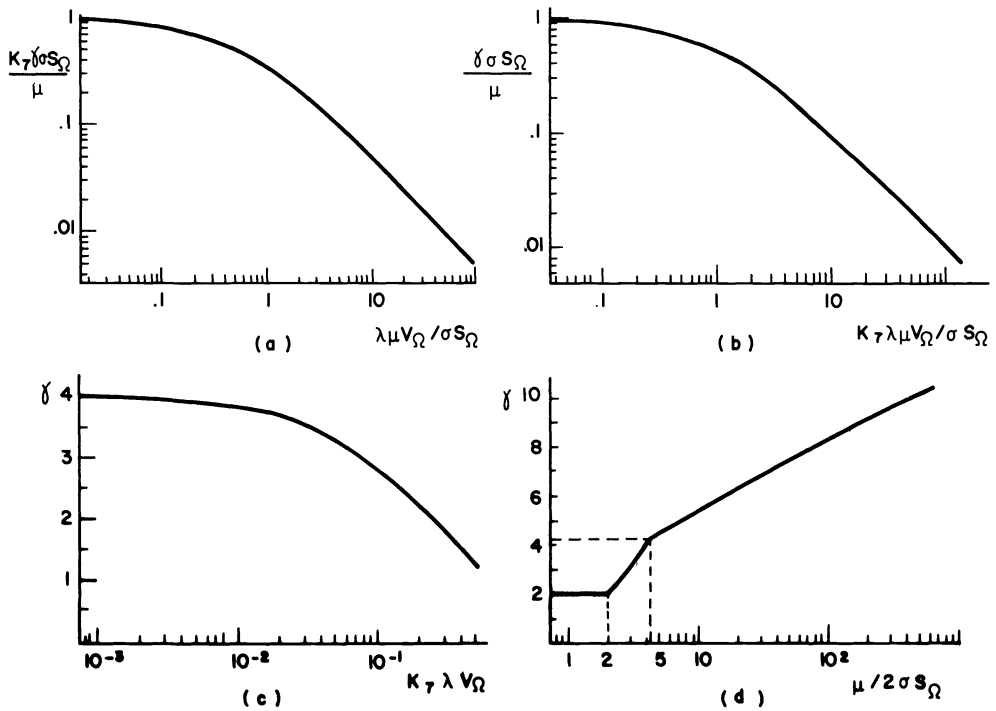


FIG. 1. *Global asymptotic stability for all* $\phi^2 > 0$. *Plots* (a)-(d) *correspond to conditions* (4.24)-(4.27).

$$\gamma \leqq 4/(1+4k_7\lambda V_\Omega), \tag{4.26}$$

$$\gamma \leqq H_1(\mu/2\sigma S_\Omega). \tag{4.27}$$

Here, k_7 is as defined in Lemma A.2, *and the positive, nondecreasing function $H_1:[0,\infty[\to$ $\mathbf{R}$ is defined by $H_1(y)=2$ for $0\leqq y\leqq 2$, $H_1(y)=y$ for $2<y\leqq y_1$, and $H_1(y)=h_2[h_1^{-1}(y)]$ for $y_1<y<\infty$, where* (i) $y_1=h_1(z_1)=4.2488\cdots$ *and $z_1=2.6761\cdots$ is the unique positive solution of the equation $z^2=\sinh z\tanh z$, and* (ii) *the strictly increasing functions, h_1, $h_2:[z_1,\infty[\to[y_1,\infty[$ are given by*

$$h_1(z)=\sinh z\tanh z/(z-\tanh z), \qquad h_2(z)=z^2/(z-\tanh z).$$

Proof. If Lemma A.2 is applied to (4.3), (4.4), one obtains

$$0\leqq u^*-u_*<1-\exp(-k_7\gamma\xi), \tag{4.28}$$

where ξ is given by (4.16) and satisfies (4.17) and

$$\begin{gathered} 1/v_*v^*\leqq 1-\xi, \quad (v_*-1)/v_*v^*<(1-\xi)^2/4, \\ (v^*-1)/(v_*-1)\geqq(1+\xi)^2/(1-\xi)^2, \end{gathered} \tag{4.29}$$

as is easily seen when taking into account (4.7).

When using (4.20), (4.21), the following inequalities are obtained from (4.5), (4.6), upon subtraction or division,

$$v^*-v_*\leqq 2\lambda\int_\Omega (u^*-u_*)\,dx+\mu^{-1}\sigma\int_{\partial\Omega}(u^*-u_*)\,ds, \tag{4.30}$$

$$v^*-v_*\leqq \lambda\int_\Omega (u^*-u_*)\,dx+\mu^{-1}\sigma[1-\exp(-\gamma\xi)]\int_{\partial\Omega}(1-u_*)\,ds, \tag{4.31}$$

$$v^*-v_*\leqq(\sigma/\mu)\left[\exp(\gamma\xi)\int_{\partial\Omega}(1-u^*)\,ds-\exp(-\gamma\xi)\int_{\partial\Omega}(1-u_*)\,ds\right], \tag{4.32}$$

$$(v^*-1)/(v_*-1)\leqq\exp(2\gamma\xi)\left[\int_{\partial\Omega}(1-u^*)\,ds\right]\Big/\left[\int_{\partial\Omega}(1-u_*)\,ds\right]. \tag{4.33}$$

(Recall that the function h identically vanishes since $g'(u)>0$ for all $0<u\leqq 1$, according to condition (A.6) of Lemma A.2.) A further substitution of (4.5), (4.20) into (4.31) yields

$$v^*-v_*\leqq\lambda\int_\Omega (u^*-u_*)\,dx+(v_*-1)[\exp(\gamma\xi)-1]. \tag{4.34}$$

When taking into account (4.7), (4.16), (4.28), (4.29), the following inequalities are obtained from (4.30)-(4.34):

$$\xi/(1-\xi)\leqq(2\lambda V_\Omega+\sigma S_\Omega/\mu)[1-\exp(-k_7\gamma\xi)], \tag{4.35}$$

$$\xi/(1-\xi)\leqq\lambda V_\Omega[1-\exp(-k_7\gamma\xi)]+(\sigma S_\Omega/\mu)[1-\exp(-\gamma\xi)], \tag{4.36}$$

$$\xi/(1-\xi)<(2\sigma S_\Omega/\mu)\sinh(\gamma\xi), \qquad (1+\xi)/(1-\xi)<\exp(\gamma\xi) \quad \text{if } \xi>0, \tag{4.37}$$

$$\xi/(1-\xi)\leqq\lambda V_\Omega[1-\exp(-\gamma k_7\xi)]+(1-\xi)[\exp(\gamma\xi)-1]/4. \tag{4.38}$$

If inequality (4.24) ((4.25) or (4.26), respectively) holds, then (4.35) ((4.36) or (4.38), respectively) yields $\xi=0$ (i.e., $v_*=v^*$); then $u_*=u^*$ (apply Lemma A.1 and take into account that $k_1=0$) and the conclusion of the theorem follows. If (4.27) holds, then $\xi=0$ and the conclusion of the theorem follows again. Use the second inequality (4.37)

if $\gamma \leqq 2$ to prove it, and observe that if $\gamma > 2$ and $\xi > 0$, then (4.27) and the first inequality (4.37) yield

$$H_1^{-1}(\gamma) < (1/\xi - 1)\sinh(\gamma\xi);$$

but this inequality cannot be satisfied for any $\xi > 0$ since the maximum of its second member, in $0 \leqq \xi \leqq 1$, is $H_1^{-1}(\gamma)$.

THEOREM 4.3 (*Global asymptotic stability for large ϕ^2*). *In addition to the assumptions of Theorem* 4.1, *let the function g satisfy conditions* (A.7) *and* (A.8) *of Lemma* A.3. *Then,* (1.1), (1.2), (2.1) *have a unique steady state, which is globally asymptotically stable for*

$$\phi^2 \geqq \phi_c^2 = \sigma(\sigma + p/\rho_1)/G(\delta), \tag{4.39}$$

if δ is such that $0 < \delta \leqq a$ and satisfies one of the following inequalities:

$$\gamma \leqq \gamma_c = (1-\delta)H_2(a_1(1-\delta)/2 + 1/2a_1(1-\delta)), \tag{4.40}$$

$$\delta \leqq \max\{(1+1/a_1)/[1+\gamma k_8(1+2a_2)],\ 2a_3/[a_4+\sqrt{a_4^2-4a_3}]\}, \tag{4.41}$$

where (i) *the strictly increasing function G and the constants ρ_1, a, and k_8 are as in Lemma* A.3; (ii) *the strictly increasing function $H_2 : [1, \infty[\to [2, \infty[$ (see Fig.* 2) *is given by $H_2(y) = 1 + y$ for $1 \leqq y \leqq 2$, $H_2(y) = h_3[h_4^{-1}(y)]$ for $2 < y < \infty$*; (iii) *the strictly increasing functions $h_3 : [0, \infty[\to [3, \infty[$ and $h_4 : [0, \infty[\to [2, \infty[$ are defined by*

$$h_3(z) = z^2 \sinh z/(z\cosh z - \sinh z), \qquad h_4(z) = (\sinh z \cosh z - z)/(z\cosh z - \sinh z);$$

and (iv) *the parameters a_1, a_2, a_3, and a_4 are*

$$a_1 = \sigma S_\Omega/\mu, \quad a_2 = \lambda\mu V_\Omega/\sigma S_\Omega, \quad a_3 = (1+1/a_1^2)/(1+a_2),$$

$$a_4 = [(1+a_1)(2+a_2) + \gamma k_8(1+2a_2)]/a_1(1+a_2).$$

Proof. If (4.39) holds and $0 < \delta \leqq a$, then (Lemma A.3)

$$0 \ll u_* \leqq u^* < \delta \leqq a, \qquad 0 \leqq u^* - u_* \leqq \delta[1 - \exp(-\gamma k_8 \xi)], \tag{4.42}$$

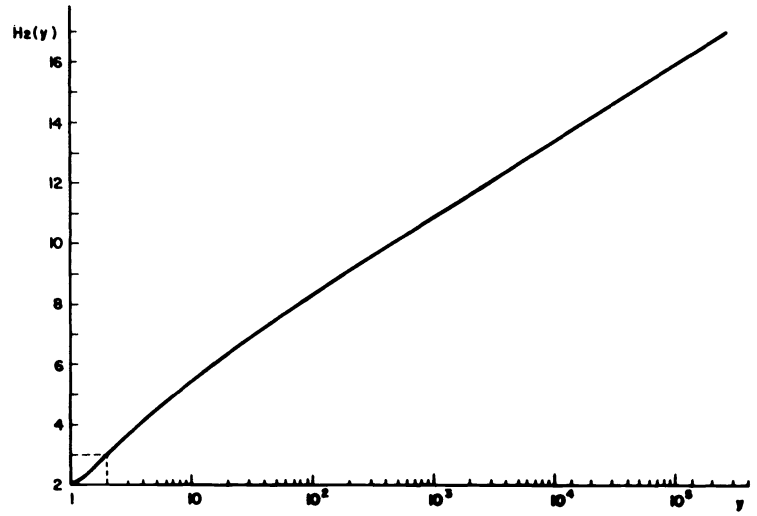


FIG. 2. *The function H_2 of Theorems* 4.3 *and* 4.4.

where $\xi = 1/v_* - 1/v^* \leqq 1$. Then if the function h is chosen to be such that $h(u) = 0$ for $0 \leqq u \leqq a$ (this may be done, with h satisfying (4.8), (4.9), since the function g satisfies (A.7)),

$$h(u_*(x)) = h(u^*(x)) = 0 \quad \text{for all } x \in \bar{\Omega}. \tag{4.43}$$

Let us first assume that ϕ^2 and δ satisfy (4.39), (4.40) and prove that $\xi = 0$. To this end, we define

$$\begin{aligned} A_* &= 1 + \mu^{-1}\sigma \exp(-\gamma\xi) \int_{\partial\Omega} (1 - u_*)\, ds, \\ A^* &= 1 + \mu^{-1}\sigma \exp(\gamma\xi) \int_{\partial\Omega} (1 - u^*)\, ds. \end{aligned} \tag{4.44}$$

$A_* \leqq v_*$ and $A^* \geqq v^*$, as it comes out from (4.5), (4.6), (4.20), (4.21), (4.43). Hence, if ξ were different from zero, it would satisfy

$$\begin{aligned} \xi = 1/v_* - 1/v^* &\leqq (A^* - A_*)/A_* A^* \\ &< 2a_1 \sinh(\gamma\xi)/[1 + a_1(1-\delta)\exp(-\gamma\xi)][1 + a_1(1-\delta)\exp(\gamma\xi)], \end{aligned} \tag{4.45}$$

or

$$1 > (1-\delta)\xi[a_1(1-\delta)/2 + 1/2a_1(1-\delta) + \cosh\gamma\xi]/\sinh\gamma\xi,$$

as obtained from (4.42), (4.44). But this inequality cannot hold for any $\xi > 0$ since the minimum of its second member, in $0 \leqq \xi < \infty$, is $(1-\delta)H_2(a_1(1-\delta)/2 + 1/2a_1(1-\delta))/\gamma$, and γ satisfies (4.40). Then, $\xi = 0$ (i.e., $v_* = v^*$), $u_* = u^*$ (Lemma A.3) and the conclusion of the theorem follows.

If ϕ^2 and δ satisfy (4.39) and (4.41), then

$$v_* \geqq \max\{1, 1 + a_1(1 - \delta - a_2\delta)\}, \qquad v^* > 1 + a_1(1-\delta), \tag{4.46}$$

as it comes out from (4.5), (4.6), (4.21), (4.42). If ξ were different from zero, (4.30), (4.42), (4.46) would yield

$$1 < a_1(1 + 2a_2)\delta[1 - \exp(-k_8\gamma\xi)]/\xi[1 + a_1(1-\delta)]\max\{1, 1 + a_1(1 - \delta - a_2\delta)\}.$$

But this inequality cannot hold for any $\xi > 0$ if δ satisfies (4.41), as is easily seen. Therefore, $\xi = 0$ and the conclusion of the theorem follows again.

Remarks. If ϕ_c^2 is calculated by means of (4.39), (4.40), then it does not depend on λ, while if it is obtained from (4.39), (4.41), then $\phi_c^2 \to \infty$ as $\lambda \to \infty$.

It is easily seen that, for fixed values of the remaining parameters, the functions $\delta \to \phi_c^2(\delta)$ and $\delta \to \gamma_c(\delta)$ are strictly decreasing in $0 < \delta < 1$. Therefore, if

$$\gamma < H_2(a_1/2 + 1/2a_1), \tag{4.47}$$

then the maximum value of δ satisfying (4.40), δ_M, is the unique solution of the equation $\gamma = \gamma_c(\delta)$. Then, the best value of ϕ_c^2 provided by (4.39), (4.40) is $\sigma(\sigma + p/\rho_1)/G(\delta)$, with $\delta = \min\{a, \delta_M\}$. If (4.47) does not hold, then (4.40) is not satisfied for any $\delta > 0$, and Theorem 4.3 does not provide a value of ϕ_c^2 uniformly valid in $0 < \lambda < \infty$. Although Theorem 4.3 provides only sufficient conditions for global asymptotic stability of the steady state, it may be seen, as a converse of Theorem 4.3 in a certain sense, that for first-order Arrhenius kinetics (i.e., for $g(u) \equiv u$) and large values of σ (see [2]), the upper linear instability bound (i.e., the supremum of the set of values of ϕ^2 such that the steady state of (1.1), (1.2), (2.1) is linearly unstable), ϕ_u^2, satisfies $\phi_u^2 \to \phi_{u0}^2 < \infty$ if $\gamma < (1 + a_1)^2/a_1$ and $\phi_u^2 \to \infty$ otherwise as $\lambda \to \infty$.

Any ϕ_c^2 provided by (4.39), (4.40), or by (4.39), (4.41), is such that $\phi_c^2 \to \infty$ as $\sigma \to \infty$. In order to calculate a value of ϕ_c^2 uniformly valid in $0 < \sigma < \infty$, which is expected to exist under mild assumptions on the function g, one would need the following result, which is stronger than that in Lemma A.3 and seemingly true (under mild assumptions on the function g): there are two constants, $\bar{\Lambda}$ and k, such that, for every $\sigma > 0$, (i) the problem (A.1) of the Appendix has a unique solution if $\Lambda \geqq \bar{\Lambda}$, and (ii) if $\bar{\Lambda} \leqq \Lambda_2 < \Lambda_1 < \infty$, then the solutions of (A.1) for $\Lambda = \Lambda_1$ and $\Lambda = \Lambda_2$, u_1 and u_2, satisfy $|u_2(x) - u_1(x)| \leqq k(\Lambda_1 - \Lambda_2)$, for all $x \in \bar{\Omega}$. Property (i) may be proved if one is able to obtain an upper multiplicity bound $\bar{\Lambda}$ when the Robin boundary data in (A.1) is replaced by Dirichlet data: $u = 1$ on $\partial\Omega$; if Ω is the unit ball of $\mathbf{R}^p$, this comes out from results by Dancer [23] that were obtained by means of topological degree theory; unfortunately, even if the results of [23] are extended to arbitrary bounded domains of $\mathbf{R}^p$, they do not seem to provide the constant k of part (ii) of the required result above. Related results in the literature, such as those in [24], [25], do not apply to our case.

Theorems 4.4 and 4.5 below provide a uniform value of ϕ_c^2 in $0 < \sigma < \infty$ but they require the function g to be strictly increasing.

THEOREM 4.4 (*Global asymptotic stability for large* ϕ^2). *In addition to the assumptions of Theorem* 4.1, *let us assume that* g *is such that* $g'(u) > 0$ *for all* $0 < u \leqq 1$, *and that*

$$2 < \gamma \leqq H_2(a_1/2 + 1/2a_1), \tag{4.48}$$

where the constant a_1 *and the function* H_2 (*see Fig.* 2) *are as defined in Theorem* 4.3. *Then,* (1.1), (1.2), (2.1) *have a unique steady state, which is globally asymptotically stable if*

$$\phi^2 \geqq 2(\mu/S_\Omega)^2 K[pS_\Omega/\mu\rho_1 + 2^{-1/p}K]/G(\delta_1), \tag{4.49}$$

where the constants δ, δ_1, *and* K *are the unique solutions of the equations*

$$\gamma = (1-\delta)H_2[a_1(1-\delta)/2 + 1/2a_1(1-\delta)], \qquad 0 < \delta < 1, \tag{4.50}$$

$$(1-\delta_1)/\sqrt{2G(\delta_1)} = (1 + D/2\rho_2)^{p-1}\sqrt{(1 + p/\sigma\rho_1)/G(\delta)}, \qquad 0 < \delta_1 < 1, \tag{4.51}$$

$$H_2(K/2 + 2/K) = \gamma a_5, \qquad K \geqq 1, \tag{4.52}$$

the constants ρ_1, ρ_2, *and* D *are as defined in Lemma* A.4, *and*

$$a_5 = (1 + D/2\rho_2)^{p-1}[p/\rho_1 + \sqrt{(p/\rho_1)^2 + 2^{1-1/p}a_6}]\sqrt{2G(1)/a_6 G(\delta_1)},$$

$$a_6 = 2(\mu/S_\Omega)^2[pS_\Omega/\mu\rho_1 + 2^{-1/p}].$$

Remarks. If $\gamma \leqq 2$, then the conclusion of the theorem is true for all $\phi^2 > 0$, according to Theorem 4.2. Equation (4.50) has a unique solution if γ satisfies (4.48), as was seen in a remark above. For a given value of δ, (4.51) has a unique solution δ_1 (which is such that $\delta_1 < \delta$), since the first member of (4.51) is a strictly decreasing function of δ_1, and it approaches 0 and ∞ as $\delta_1 \to 1$ and as $\delta_1 \to 0$, respectively. Since the second member of (4.52) is larger than 2 ($a_5 > 1$ and $\gamma > 2$), (4.52) has a unique solution (recall that $H_2(1) = 2$, H_2 is strictly increasing and $H_2(y) \to \infty$ as $y \to \infty$).

Proof of Theorem 4.4. If $\phi^2 G(\delta) \exp(\gamma - \gamma/v_*) \geqq \sigma(\sigma + p/\rho_1)$, then $u_* \leqq u^* \leqq \delta$ (Lemma A.3) and, as in the proof of Theorem 4.3, $\xi = 1/v_* - 1/v^*$ is seen to satisfy (4.45), which implies $\xi = 0$ (i.e., $v_* = v^*$). Then $u_* = u^*$ (Lemma A.1), and the conclusion of the theorem follows.

If $\phi^2 G(\delta) \exp(\gamma - \gamma/v_*) < \sigma(\sigma + p/\rho_1)$, then $u_m^* = \min\{u^*(x): x \in \partial\Omega\}$ satisfy (Lemma A.4)

$$(1 - u_m^*)/\sqrt{2G(u_m^*)} < (1 + D/2\rho_2)^{p-1}\sqrt{(1 + p/\sigma\rho_1)/G(\delta)}. \tag{4.53}$$

Since the first member of (4.53) is a strictly decreasing function of u_m^* and δ_1 satisfies (4.51), $\delta_1 \leqq u_m^*$ and

$$\delta_1 < u^*(x) \quad \text{for all } x \in \partial\Omega. \tag{4.54}$$

Then, u_m^* and $u_M^* = \max\{u^*(x): x \in \partial\Omega\} = \max\{u^*(x): x \in \bar{\Omega}\}$ satisfy (Lemma A.4)

$$\sigma(1-u_m^*) < (1+D/2\rho_2)^{p-1}\sqrt{2\Lambda_* G(1)}, \tag{4.55}$$

$$\sigma(1-u_M^*) > \Lambda_* G(\delta_1)/[p/\rho_1 + \sqrt{(p/\rho_1)^2 + 2^{1-1/p}\Lambda_* G(\delta_1)}] = \mu K_*/S_\Omega \tag{4.56}$$

where

$$\Lambda_* = \phi^2 \exp(\gamma - \gamma/v_*). \tag{4.57}$$

Since $K \geqq 1$ and ϕ^2 satisfies (4.49), we have $\Lambda_* G(\delta_1) \geqq \phi^2 G(\delta_1) \geqq a_6$, and

$$(1+D/2\rho_2)^{p-1}\sqrt{2\Lambda_* G(1)}\, S_\Omega/\mu K_* \leqq a_5. \tag{4.58}$$

Furthermore, (4.49), (4.56), (4.57) yield

$$K \leqq K_*. \tag{4.59}$$

Then, if A_* and A^* are as defined by (4.44), $\xi = 1/v_* - 1/v^* \leqq (A^* - A_*)/A_* A^*$ must vanish because otherwise it would satisfy

$$\xi < a_5 \sinh(\gamma\xi)/(K/2 + 1/2K + \cosh\gamma\xi),$$

as it comes out from (4.44), (4.55), (4.56), (4.58), (4.59), or

$$a_5 > (K/2 + 1/2K + \cosh\gamma\xi)/\sinh\gamma\xi,$$

and this inequality cannot hold for any $\xi > 0$ since the minimum of its second member in $0 \leqq \xi < \infty$ is $H_2(K/2 + 1/2K)/\gamma$, and K satisfies (4.52). Therefore, $\xi = 0$ (i.e., $v_* = v^*$), $u_* = u^*$ (Lemma A.1) and the conclusion of the theorem follows.

Observe that the second member of (4.49) does not depend on λ. The following theorem provides a better result if λ is sufficiently small. It also applies for arbitrarily large values of σ.

THEOREM 4.5 (*Global asymptotic stability for large* ϕ^2). *In addition to the hypothesis of Theorem* 4.1, *let us assume that the function g satisfies condition* (A.6) *of Lemma* A.2, *and that*

$$\gamma > 2, \quad \delta = \max\{(1+1/a_1)/[1+\gamma k_7(1+2a_2)],\ 2a_3/[a_7 + \sqrt{a_7^2 - 4a_3}]\} < 1, \tag{4.60}$$

where the constants a_1, a_2, *and* a_3 *are as in Theorem* 4.3, k_7 *is as defined in Lemma* A.2, *and*

$$a_7 = [(1+a_1)(2+a_2) + \gamma k_7(1+2a_2)]/a_1(1+a_2).$$

Then, (1.1), (1.2), (2.1) *possess a unique solution, which is globally asymptotically stable if*

$$\phi^2 \geqq 2(\mu/S_\Omega)^2 K[pS_\Omega/\mu\rho_1 + 2^{-1/p}K]/G(\delta_1),$$

where δ_1 *is the unique solution of*

$$(1-\delta_1)/\sqrt{2G(\delta_1)} = (1+D/2\rho_2)^{p-1}\sqrt{(1+p/\sigma\rho_1)/G(\delta)}, \qquad 0 < \delta_1 < 1,$$

the constants ρ_1, ρ_2, *and* D *are as in Lemma* A.4, *and*

$$K = [\gamma - 2 + \lambda V_\Omega + \sqrt{(\gamma + \lambda V_\Omega)^2 + 4\gamma\lambda k_7 V_\Omega}]/2. \tag{4.61}$$

Remark. If $\gamma \leqq 2$ or if $\delta \geqq 1$, then the conclusion of the theorem is true for all $\phi^2 > 0$, according to Theorem 4.2.

Proof. If $\phi^2 G(\delta)\exp(\gamma-\gamma/v_*)\geqq\sigma(\sigma+p/\rho_1)$, then $u_*\leqq u^*\leqq\delta$ (Lemma A.3) and, as in the proof of Theorem 4.3, $\xi=1/v_*-1/v^*$ is seen to satisfy

$$\xi\leqq a_1(1+2a_2)\delta[1-\exp(-k_7\gamma\xi)]/[1+a_1(1-\delta)]\max\{1,1+a_1(1-\delta-a_2\delta)\}.$$

This inequality cannot hold for any $\xi>0$ if δ is given by (4.60). Therefore, $\xi=0$ (i.e., $v_*=v^*$), $u_*=u^*$ (Lemma A.1) and the conclusion of the theorem follows.

If $\phi^2 G(\delta)\exp(\gamma-\gamma/v_*)<\sigma(\sigma+p/\rho_1)$, then $u_m^*=\min\{u^*(x): x\in\partial\Omega\}$ satisfies (4.53) (Lemma A.4). As in the proof of Theorem 4.4, this implies that u^* satisfies (4.54). Then u_m^* and $u_M^*=\max\{u^*(x): x\in\partial\Omega\}$ are seen to satisfy (4.55), (4.56), where Λ_* is given again by (4.57), and K satisfies (4.59) again. In addition, u_* and u^* satisfy (4.28) (Lemma A.2). Then B_* and B^*, which are defined by

$$B_*=1-\lambda\int_\Omega(u^*-u_*)\,dx+\mu^{-1}\sigma\int_{\partial\Omega}(1-u^*)\,ds,$$

$$B^*=1+\mu^{-1}\sigma\exp(\gamma\xi)\int_{\partial\Omega}(1-u^*)\,ds,$$

satisfy

$$0\leqq B^*-B_*\leqq\lambda\int_\Omega(u^*-u_*)\,dx+(B^*-1)[1-\exp(-\gamma\xi)],$$
$$B_*\geqq 1+K-\lambda V_\Omega,\quad B^*\geqq 1+K, \tag{4.62}$$

as it comes out from (4.56), (4.59). Also, $B_*\leqq v_*$ and $B^*\geqq v^*$ (see (4.5), (4.6), (4.21)). Hence, $\xi=1/v_*-1/v^*\leqq(B^*-B_*)/B_*B^*$ satisfies

$$\xi\leqq[1-\exp(-\gamma\xi)]/(1+K-\lambda V_\Omega)+\lambda V_\Omega[1-\exp(-k_7\gamma\xi)]/(1+K)(1+K-\lambda V_\Omega),$$

as obtained from (4.28), (4.62). But this inequality cannot hold for any $\xi>0$ if K is given by (4.61), as it is easily seen. Therefore $\xi=0$ and the conclusion of the theorem follows.

Finally, since the steady-state solutions of (1.1), (1.2) do not depend on the parameter λ, the following corollary is true.

COROLLARY 4.6. *If, for some $\lambda>0$, the hypothesis of one of the Theorems* 4.1-4.5 *hold, then* (1.1), (1.2) *has a unique steady state.*

5. Concluding remarks. A sequence of nested, globally finitely attracting, invariant regions of the phase space of (1.1), (1.2), (2.1), converging to an invariant, globally asymptotically stable region, has been obtained in § 3. In § 4, some quantitative sufficient conditions (ϕ^2 sufficiently large or small, or g increasing and γ sufficiently small) for global asymptotic stability of the steady state have been obtained, for a kinetic function f of the type $f(u,v)=g(u)\exp(\gamma-\gamma/v)$. Some of the results, which were not uniformly valid in $0<\lambda<\infty$ if γ is too large, have been explained by comparison with linear stability results that were obtained in [2]. Of course, similar results to those of § 4 may be obtained for any kinetic function satisfying assumption (H.2), such as that in (1.6).

The results of § 3 remain valid when the Robin type of boundary data is replaced by Dirichlet boundary data ($u=1$ on $\partial\Omega$), and $\sigma\int_{\partial\Omega}(1-u)\,ds$ is replaced by $\int_{\partial\Omega}(\partial u/\partial n)\,ds$ everywhere. To see that, a unit order (see, e.g., Amann [26]) must be used to replace the definition of the order relation $\ll$ at the end of § 1 by $u_1\ll u_2$ means that there is a positive constant c such that $u_1(x)+ce(x)\leqq u_2(x)$ for all $x\in\Omega$, where the unit e is defined by $\Delta e+1=0$ in Ω, $e=0$ on $\partial\Omega$. Such order definition could have been used in § 3 to obtain results for both Robin and Dirichlet problems at the same time, although it has not been done for the sake of clarity.

Growth restrictions on the function f are not necessary for the ideas of § 3 to apply. The assumption $f_v(u, v) > 0$ has been imposed because it is satisfied by the most commonly used kinetic functions (i.e., by those in (1.4)-(1.6)), but it could be removed; then the definition of the sequence (3.1)-(3.7) should be changed somewhat.

The ideas of this paper are naturally extended if: (a) the Laplacian operator Δ is replaced by a uniformly strongly elliptic operator; (b) the function f and/or the boundary data depend on the space variable x; or (c) the linear boundary conditions in (1.1) are replaced by appropriate nonlinear ones. They apply also to some more general reaction-diffusion problems, such as the nonisothermal model (1.1), (1.3) (this point is currently under research). Nevertheless, (1.1), (1.2) has been considered first because such isothermal model (a) has practical interest in itself (not only as a limit of (1.1), (1.3)), as was explained in the Introduction, and (b) it retains the main intrinsic difficulty of (1.1), (1.3), namely, the flow defined by (1.1), (1.2) is not monotone. Also, global stability results for (1.1)-(1.2) may be (and have been) compared with local stability results, which were obtained in [2] for the slab geometry and first-order Arrhenius kinetics.

Appendix. Let us consider the elliptic semilinear problem

$$\Delta u = \Lambda g(u) \quad \text{in } \Omega, \qquad \partial u/\partial n = \sigma(1-u) \quad \text{on } \partial\Omega, \tag{A.1}$$

where $\Lambda \geqq 0$, $\sigma > 0$, $\Omega \subset \mathbf{R}^p$ ($p = 1$, 2, or 3) satisfies assumption (H.1) and the C^1-function g satisfies (4.2).

LEMMA A.1. *Under the assumptions above*:

A. *The problem* (A.1) *possesses a minimal and a maximal solution,* $\underline{u}$, $\tilde{u} \in C^2(\bar{\Omega})$ *such that*

$$0 \ll \underline{u} \leqq \tilde{u} \ll 1. \tag{A.2}$$

B. *The solution of* (A.1) *is unique if* $0 \leqq \Lambda < \alpha/k_1 \leqq \infty$, *where* $\alpha > 0$ *is the smallest eigenvalue of* (2.2), *and* k_1 *is given by* (4.9). *Furthermore, if* u_1 *and* u_2 *are the solutions of* (A.1) *for* $\Lambda = \Lambda_1$ *and for* $\Lambda = \Lambda_2$ *with* $0 \leqq \Lambda_2 < \Lambda_1 < \alpha/k_1 \leqq \infty$, *then*

$$0 \ll u_2 - u_1 \leqq k_2[k_3 + k_4\Lambda_1 k_1/(\alpha - \Lambda_1 k_1)](\Lambda_1 - \Lambda_2), \tag{A.3}$$

where

$$k_2 = \max\{g(u): 0 \leqq u \leqq 1\}, \quad k_3 = \max\{\psi_1(x): x \in \bar{\Omega}\}, \quad k_4 = \max\{\psi_2(x): x \in \bar{\Omega}\},$$

$\psi_1 \gg 0$ *is the unique solution of*

$$\Delta\psi_1 + 1 = 0 \quad \text{in } \Omega, \qquad \partial\psi_1/\partial n + \sigma\psi_1 = 0 \quad \text{on } \partial\Omega,$$

and ψ_2 *is any eigenfunction of* (2.2) *such that* $\psi_2 \geqq \psi_1$.

Proof. A. For the existence of the minimal and maximal solutions of (A.1) see, e.g., [13], [14], or [20]. Inequalities (A.2) follow by standard arguments, using maximum principles.

B. Since the function $u \to g(u) + k_1 u$ is nondecreasing in $0 \leqq u \leqq 1$, $U = \tilde{u} - \underline{u}$ satisfies

$$\Delta U + \Lambda k_1 U \geqq 0 \quad \text{in } \Omega, \qquad \partial U/\partial n + \sigma U = 0 \quad \text{on } \partial\Omega. \tag{A.4}$$

Then, if $\Lambda k_1 < \alpha$, the generalized maximum principle (see [17]) shows that $U \leqq 0$. Therefore, $\underline{u} = \tilde{u}$ and the solution of (A.1) is unique.

Since the smallest eigenvalue of (2.2) depends continuously on the parameter σ (see, e.g., [22]), one may choose $\varepsilon > 0$ sufficiently small for the smallest eigenvalue of

$$\Delta\psi + \alpha_1\psi = 0 \quad \text{in } \Omega, \qquad \partial\psi/\partial n + (\sigma - \varepsilon)\psi = 0 \quad \text{on } \partial\Omega, \tag{A.5}$$

α_1, to be such that $k_1\Lambda_1 \leqq \alpha_1$. Then if $\psi \gg 0$ is an eigenfunction of (A.5), $U = (u_2 - u_1)/\psi$ satisfies

$$\psi \Delta U + 2\nabla\psi \cdot \nabla U \leqq \Lambda_1[g(u_2) - g(u_1)] + \alpha_1(u_2 - u_1) \quad \text{in } \Omega, \qquad \partial U/\partial n + \varepsilon U = 0 \quad \text{on } \partial\Omega.$$

Then, standard maximum principles show that $U \gg 0$, i.e., that $u_2 \gg u_1$.

Finally, $U = u_2 - u_1 - (\Lambda_1 - \Lambda_2)k_2[\psi_1 + \Lambda_1 k_1 \psi_2/(\alpha - \Lambda_1 k_1)]$ is easily seen to satisfy (A.4) with $\Lambda = \Lambda_1$. Therefore, $U \leqq 0$ and the second inequality (A.3) readily follows.

Remark. If $g'(u) \geqq 0$ for all $u \in [0, 1]$, then $k_1 = 0$, the solution of (A.1) is unique for all $\Lambda \geqq 0$ and inequalities (A.3) become $0 \ll u_2 - u_1 \leqq k_2 k_3(\Lambda_1 - \Lambda_2)$. Under an additional mild assumption on the function g, the following lemma provides another upper bound to $u_2 - u_1$, which is stronger than that above when Λ_2 is large.

LEMMA A.2. *In addition to the assumptions of Lemma* A.1, *let us assume that*

$$k_7 = \sup\{g(u)/ug'(u) : 0 < u \leqq 1\} < \infty. \tag{A.6}$$

Let u_1 and u_2 be the solutions of (A.1) *for $\Lambda = \Lambda_1$ and $\Lambda = \Lambda_2$ with $0 \leqq \Lambda_2 < \Lambda_1 < \infty$. Then*

$$0 \ll u_2 - u_1 \leqq [1 - (\Lambda_2/\Lambda_1)^{k_7}] \max\{u_2(x) : x \in \bar{\Omega}\}.$$

Remark. Assumption (A.6) implies that $g'(u) > 0$ for all $0 < u \leqq 1$. Although the converse is not true in general, it is true if, for example, the function $u \to g'(u)$ is nondecreasing in a neighborhood of $u = 0$, as is the case for most commonly used kinetic functions (e.g., for those given in (1.4), (1.5)).

Proof. $U = u_2 - u_1 \gg 0$ satisfies

$$\Delta U = \Lambda_2 g(u_2) - \Lambda_1 g(u_1) \quad \text{in } \Omega, \qquad \partial U/\partial n + \sigma U = 0 \quad \text{on } \partial\Omega.$$

Let x_0 be a point (not necessarily unique) where the maximum of U is attained. Since $\sigma > 0$ and $U(x_0) > 0$, x_0 cannot be a point of $\partial\Omega$. Then, $\Delta U \leqq 0$ at $x = x_0$ and

$$\Lambda_2/\Lambda_1 \leqq g(u_1(x_0))/g(u_2(x_0)) \leqq [u_1(x_0)/u_2(x_0)]^{1/k_7},$$

where the second inequality is easily obtained when using (A.6) (the function $u \to g(u)/u^{1/k_7}$ is nondecreasing). Then the conclusion of the lemma readily follows.

Let us assume now that the function g is such that there exists a constant a, $0 < a \leqq 1$, satisfying

$$g'(u) > 0 \quad \text{for all } 0 < u \leqq a, \qquad g(a) < g(u) \quad \text{for all } a < u \leqq 1, \tag{A.7}$$

$$k_8 = \sup\{g(u)/ug'(u) : 0 < u \leqq a\} < \infty. \tag{A.8}$$

Then, we have the following.

LEMMA A.3. *Let us assume that, in addition to the hypothesis of Lemma* A.1, (A.7) *holds. If*

$$\Lambda G(a) \geqq \sigma(\sigma + p/\rho_1) \tag{A.9}$$

then (A.1) *has a unique solution, $u = u(x)$, which satisfies*

$$\Lambda G(u(x)) \leqq \sigma(\sigma + p/\rho_1) \quad \textit{for all } x \in \bar{\Omega}, \tag{A.10}$$

where ρ_1 is defined in the interior sphere property (*assumption* (H.1); *see Introduction*), *and $G : [0, 1] \to \mathbf{R}$ is the strictly increasing function*

$$G(u) = \int_0^u g(z)\, dz. \tag{A.11}$$

If, in addition, (A.8) *holds and if* u_1 *and* u_2 *are the solutions of* (A.1) *for* $\Lambda=\Lambda_1$ *and* $\Lambda=\Lambda_2$, *where* Λ_1 *and* Λ_2 *satisfy* (A.9) *and* $\Lambda_2<\Lambda_1<\infty$, *then*

$$0 \ll u_2-u_1 \leqq [1-(\Lambda_2/\Lambda_1)^{k_8}] \max\{u_2(x): x\in\bar{\Omega}\}. \tag{A.12}$$

Proof. Let $u=u(x)$ be a solution of (A.1) and let x_0 be a point (not necessarily unique) of $\bar{\Omega}$ where the maximum of u, $u_M=\max\{u(x): x\in\bar{\Omega}\}$, is attained. $x_0\in\partial\Omega$ because otherwise $\Delta u(x_0)>0$. Let $S_1\subset\Omega$ be the hypersphere, of radius ρ_1, that is tangent to $\partial\Omega$ at x_0. We consider the problem

$$\Delta w=\Lambda g_1(w) \quad \text{in } S_1, \qquad w=u_M \quad \text{on } \partial S_1, \tag{A.13}$$

where the C^1-function $g_1:[0,1]\to\mathbf{R}$ is such that $g_1(u)=g(u)$ for $0\leqq u\leqq a$, $g_1'(u)>0$ for $a<u\leqq 1$. Problem (A.13) has a unique solution (Lemma A.1), which is spherically symmetric (Gidas et al. [27]), and given by

$$r^{1-p}\,d[r^{p-1}\,dw/dr]/dr\equiv d^2w/dr^2+(p-1)r^{-1}\,dw/dr=\Lambda g_1(w) \quad \text{in } 0<r<\rho_1, \tag{A.14}$$

$$dw/dr=0 \quad \text{at } r=0, \qquad w=u_M \quad \text{at } r=\rho_1, \tag{A.15}$$

where $r=\overline{x_1x}$ and x_1 is the center of S_1. Furthermore, the solution of (A.13) satisfies $w(x)\geqq u(x)$ for all $x\in S_1$, as it is easily seen by means of maximum principles. Hence

$$\sigma(1-u_M)=(\partial u/\partial n)_{x=x_0}\geqq(dw/dr)_{r=\rho_1}. \tag{A.16}$$

On the other hand, integration of (A.14), (A.15) yields

$$r^{p-1}\,dw/dr=\Lambda\int_0^r z^{p-1}g_1(w(z))\,dz. \tag{A.17}$$

Therefore, the function $r\to w(r)$ is strictly increasing and (A.17) yields $p\,dw/dr<\Lambda r g_1(w(r))$ for all $0<r\leqq\rho_1$. Hence, (A.14) implies that the function $r\to dw/dr$ is also strictly increasing, and (A.17) leads to

$$\begin{aligned}\rho_1^{p-1}(dw/dr)^2_{r=\rho_1} &> \Lambda\int_{\varepsilon\rho_1}^{\rho_1} z^{p-1}g_1(w(z))(dw/dz)\,dz\\ &> \Lambda(\varepsilon\rho_1)^{p-1}[G_1(u_M)-G_1(w(\varepsilon\rho_1))],\end{aligned} \tag{A.18}$$

for any real constant ε such that $0<\varepsilon<1$, where

$$G_1(u)=\int_0^u g_1(z)\,dz. \tag{A.19}$$

But, as it is seen from (A.17), (A.19),

$$\rho_1^{p-1}(dw/dr)_{r=\rho_1}>\Lambda g_1(w(\varepsilon\rho_1))\int_{\varepsilon\rho_1}^{\rho_1} z^{p-1}\,dz=\Lambda g_1(w(\varepsilon\rho_1))\rho_1^p(1-\varepsilon^p)/p, \tag{A.20}$$

$$G_1(w(\varepsilon\rho_1))<w(\varepsilon\rho_1)g_1(w(\varepsilon\rho_1))<u_M g_1(w(\varepsilon\rho_1)). \tag{A.21}$$

Equations (A.16), (A.18), (A.20)-(A.21) lead to the inequality

$$\Lambda G_1(u_M)<\sigma(1-u_M)[\varepsilon^{1-p}\sigma(1-u_M)+pu_M/\rho_1(1-\varepsilon^p)],$$

which is valid for $0<\varepsilon<1$. Then, when replacing ε^p by $1-u_M$, we obtain (recall that $0<u_M<1$)

$$\Lambda G_1(u_M)<\sigma(\sigma+p/\rho_1). \tag{A.22}$$

Since the function G_1 is strictly increasing and $G_1(u)=G(u)$ for $0\leqq u\leqq a$, if Λ satisfies (A.9), then (A.22) yields $u_M<a$, i.e., any solution of (A.1) satisfies

$$u(x)<a \quad \text{for all } x\in\bar{\Omega}. \tag{A.23}$$

Then, (A.1) has a unique solution, as it comes out when Lemma A.1 and maximum principles are applied and (A.7) and (A.23) are taken into account. Inequality (A.10) is readily obtained from (A.22).

Finally, (A.12) is obtained by the argument of the proof of Lemma A.2, when taking into account that u_1 and u_2 satisfy (A.23).

LEMMA A.4. *In addition to the assumptions of Lemma* A.1, *let us assume that* $g'(u)>0$ *for all* $0<u\leqq 1$, *and let* $u=u(x)$ *be the (unique) solution of* (A.1) *for a given value of* $\Lambda>0$. *Then,* $u_M=\max\{u(x): x\in\partial\Omega\}=\max\{u(x): x\in\bar{\Omega}\}$, and $u_m=\min\{u(x): x\in\partial\Omega\}$ *satisfy*

$$\sigma(1-u_M)>\Lambda G(u_M)/[p/\rho_1+\sqrt{(p/\rho_1)^2+2^{1-1/p}\Lambda G(u_M)}], \tag{A.24}$$

$$\sigma(1-u_m)<(1+D/2\rho_2)^{p-1}\sqrt{2\Lambda G(u_m)}, \tag{A.25}$$

where ρ_1 *and* ρ_2 *are defined in the interior and exterior sphere properties (assumption* (H.1); *see Introduction), D is the diameter of* Ω *and the strictly increasing function* $G:[0,1]\to\mathbf{R}$ *is defined by* (A.11).

Proof. The argument that led to (A.21) in the proof of Lemma A.3 shows that u_M satisfies

$$\Lambda G(u_M)<\sigma(1-u_M)[\varepsilon^{1-p}\sigma(1-u_M)+pu_M/\rho_1(1-\varepsilon^p)]$$

for all $0<\varepsilon<1$. Then, if $\varepsilon^p=\frac{1}{2}$ (A.24) is readily obtained.

Let $x_0\in\partial\Omega$ be a point (not necessarily unique) where u_m is attained. Let S_2 be the hypersphere of radius ρ_2, tangent to $\partial\Omega$ at x_0 to which the exterior sphere property refers. Let x_2 be the center of S_2 and let S be the hypersphere of center at x_2 radius ρ_2+D. Then $\Omega\subset S-\bar{S}_2$. Let $w:\bar{S}-S_2\to\mathbf{R}$ be defined by

$$\Delta w=\Lambda g(w) \quad \text{in } S-\bar{S}_2, \qquad w=u_m \quad \text{on } \partial(S-\bar{S}_2). \tag{A.26}$$

Problem (A.26) possesses a unique solution, which is spherically symmetric, and given by

$$\begin{aligned} r^{1-p}\,d[r^{p-1}\,dw/dr]/dr&=\Lambda g(w) \quad \text{in } \rho_2<r<\rho_2+D,\\ w&=u_m \quad \text{at } r=\rho_2 \text{ and } r=\rho_2+D \end{aligned} \tag{A.27}$$

where $r=\overline{x_2x}$. To see this, observe that (A.27) has (at least) a solution (see, e.g., Keller [13]), and that (A.26) has (at most) one solution (Lemma A.1).

The (unique) solution of (A.26) satisfies (apply maximum principles)

$$0<w(x)<u_m \quad \text{for all } x\in S-\bar{S}_2, \qquad w(x)\leqq u(x) \quad \text{for all } x\in\bar{\Omega}.$$

Hence

$$-(dw/dr)_{r=\rho_2}\geqq(\partial u/\partial n)_{x=x_0}=\sigma(1-u_m). \tag{A.28}$$

On the other hand, let $r_1>\rho_2$ be the smallest value of r where $dw/dr=0$. Since $r_1\leqq\rho_2+D/2$ (see Gidas et al. [27]), when (A.27) is multiplied by $r^{2p-2}\,dw/dr$ and the resulting equation is integrated between ρ_2 and r_1, we obtain

$$\begin{aligned} [\rho_2^{p-1}(dw/dr)_{r=\rho_2}]^2&=-2\Lambda\int_{\rho_2}^{r_1}r^{2p-2}g(w(r))(dw/dr)\,dr\\ &\leqq 2\Lambda r_1^{2p-2}\int_{w_1}^{u_m}g(w)\,dw<2\Lambda(\rho_2+D/2)^{2p-2}G(u_m) \end{aligned} \tag{A.29}$$

where $w_1=w(r_1)>0$. Then (A.25) readily follows from (A.28), (A.29), taking into account that $(dw/dr)_{r=\rho_2}<0$.

REFERENCES

[1] R. ARIS, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts, Vol.* I *and* II, Clarendon Press, Oxford, 1975.

[2] I. E. PARRA AND J. M. VEGA, *Local nonlinear stability of the steady state in an isothermal catalyst*, SIAM J. Appl. Math., 48 (1988), to appear.

[3] C. J. PEREIRA AND A. VARMA, *Uniqueness criteria of the steady state in automotive catalysis*, Chem. Engrg. Sci., 33 (1978), pp. 1645-1657.

[4] H. AMANN, *Existence and stability of solutions for semilinear parabolic systems, and applications to some diffusion-reaction equations*, Proc. Roy. Soc. Edinburgh, 81A (1978), pp. 35-47.

[5] C. V. PAO, *Asymptotic stability of reaction-diffusion systems in chemical reactor and combustion theory*, J. Math. Anal. Appl., 82 (1981), pp. 503-526.

[6] J. HERNÁNDEZ, *Some existence and stability results for solutions of reaction-diffusion systems with nonlinear boundary conditions*, in Nonlinear Differential Equations: Invariance Stability and Bifurcation, P. de Mottoni, ed., Academic Press, New York, 1981.

[7] A. LEUNG AND D. CLARK, *Bifurcations and large-time asymptotic behavior for prey-predator reaction-diffusion equations with Dirichlet boundary data*, J. Differential Equations, 35 (1980) pp. 113-127.

[8] C. V. PAO, *On nonlinear reaction-diffusion systems*, J. Math. Anal. Appl., 87 (1982), pp. 165-198.

[9] ———, *Asymptotic stability of a coupled diffusion system arising from gas-liquid reactions*, Rocky Mountain J. Math., 12 (1982), pp. 55-73.

[10] G. S. LADDE, V. LAKSHMIKANTHAM, AND A. S. VATSALA, *Existence and asymptotic behavior of reaction-diffusion systems via coupled quasi-solutions*, in Nonlinear Analysis and Applications, V. Lakshmikantham, ed., Academic Press, New York, 1982.

[11] M. W. HIRSCH, *The dynamical systems approach to differential equations*, Bull. Amer. Math. Soc., 11 (1984), pp. 1-64.

[12] A. LEUNG, *Monotone schemes for semilinear elliptic systems related to ecology*, Math. Meth. Appl. Sci., 4 (1982), pp. 272-285.

[13] H. B. KELLER, *Elliptic boundary value problems suggested by nonlinear diffusion processes*, Arch. Rational Mech. Anal., 35 (1969), pp. 363-381.

[14] D. H. SATTINGER, *Monotone methods in nonlinear elliptic and parabolic boundary value problems*, Indiana Univ. Math. J., 21 (1972), pp. 979-1000.

[15] J. C. BURNELL, A. A. LACEY, AND G. C. WAKE, *Steady states of the reaction-diffusion equations. Part* I: *Questions of existence and continuity of solution branches*, J. Austral. Math. Soc. Ser. B, 24 (1983), pp. 374-391.

[16] ———, *Steady states of the reaction-diffusion equations. Part* II: *Uniqueness of solutions and some special cases*, J. Austral. Math. Soc. Ser. B, 24 (1983), pp. 392-416.

[17] M. H. PROTTER AND H. F. WEINBERGER, *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1967.

[18] D. HENRY, *Geometric theory of semilinear parabolic equations*, in Lecture Notes in Math. 840, Springer-Verlag, New York-Berlin, 1981.

[19] H. AMANN, *Periodic solutions of semilinear parabolic equations*, in Nonlinear Analysis: A Collection of Papers in Honour of Erich H. Rothe, L. Cesary, R. Kannan, and H. F. Weinberger, eds., Academic Press, New York, 1978.

[20] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York-Berlin, 1983.

[21] W. H. RAY AND S. P. HASTINGS, *The influence of the Lewis number on the dynamics of chemically reacting systems*, Chem. Engrg. Sci., 35 (1980), pp. 589-595.

[22] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics, Vol.* I, Wiley-Interscience, New York, 1953.

[23] E. N. DANCER, *On the structure of solutions of an equation in catalysis theory when a parameter is large*, J. Differential Equations, 37 (1980), pp. 404-437.

[24] A. CASTRO AND R. SHIVAJI, *Uniqueness of positive solutions for a class of elliptic boundary value problems*, Proc. Roy. Soc. Edinburgh, 98A (1984), pp. 267-269.

[25] R. SHIVAJI, *Remarks on an S-shaped bifurcation curve*, J. Math. Anal. Appl., to appear.

[26] H. AMANN, *Fixed point equations and nonlinear eigenvalue problems in ordered Banach spaces*, SIAM Rev., 18 (1976), pp. 620-708.

[27] B. GIDAS, W. NI, AND L. NIRENBERG, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209-243.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

003

TRANSPORT EQUATIONS WITH SECOND-ORDER DIFFERENTIAL COLLISION OPERATORS*

CHRIS COSNER†, SUZANNE M. LENHART‡, AND VLADIMIR PROTOPOPESCU§

Abstract. This paper discusses existence, uniqueness, and a priori estimates for time-dependent and time-independent transport equations with unbounded collision operators. These collision operators are described by second-order differential operators resulting from diffusion in the velocity space. The transport equations are degenerate parabolic-elliptic partial differential equations, that are treated by modifications of the Fichera–Oleinik–Radkevic Theory of second-order equations with nonnegative characteristic form. We consider weak solutions in spaces that are extensions of L^p to include traces on certain parts of the boundary. This extension is necessary due to the nonclassical boundary conditions imposed by the transport problem, which requires a specific analysis of the behavior of our weak solutions.

Key words. transport equations, electron scattering, degenerate parabolic-elliptic PDE, traces

AMS(MOS) subject classifications. 35K65, 82A70

1. Introduction. Recently, an abstract theory of the time-dependent transport equations has been given [4], covering a fairly large number of possible applications. The transport equations in [4] are first-order partial differential equations, with general time-dependent phase spaces, boundary conditions, and transport operators (irrespective of dimensionality), and they are analyzed largely by the method of characteristics. In the present article, we consider transport equations with unbounded collision operators, that are described by second-order differential operators resulting from diffusion in the velocity space. Clearly the method of characteristics does not work for these equations which are of degenerate parabolic-elliptic type. Instead we use the theory of second-order equations with nonnegative characteristic form developed by Fichera and Oleinik and Radkevic [10]. Our analysis extends that of [10] by giving a more careful discussion of boundary values of weak solutions. This extension is necessary due to the nonclassical boundary conditions imposed by the transport problem.

The results of [4] are quite general. The major limitation of the theory presented in [4] is related to the collision part of the transport operator. Namely, (i) the collision operator is assumed to be separable into a sum of two operators describing the "in" and "out" scattering, respectively, and (ii) the "out" operator is required to be a bounded operator. The second assumption allows them to consider the transport operators as a bounded perturbation of a first-order differential operator derived from a real vector field. Accordingly, the strategy to obtain the existence and uniqueness results in [4] was to use the method of characteristics for the vector field and a perturbative approach for the full transport operator. Obviously, this strategy fails when the collision operator is not bounded with respect to the vector field. This happens to be the case in several important physical problems such as Brownian motion [6]

* Received by the editors October 15, 1986; accepted for publication (in revised form) June 25, 1987.

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

‡ Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1300. The work of this author was supported in part by the Institute for Mathematics and its Applications (with funds from the National Science Foundation and the Army Research Office), National Science Foundation grant DMS-8508651, and University of Tennessee Faculty Leave Award and Science Alliance Award.

§ Engineering Physics and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831. The work of this author was supported in part by Department of Energy contract DE-AC05-840R21400, and by the Center for Studies of Nonlinear Phenomena at Oak Ridge National Laboratory.

and electron scattering [5], in which collisions are characterized by very small deflection angles. Mathematically, these collisions are described by second-order differential operators accounting for diffusion in the velocity space. Since the collision operator acts only on the velocity variable, this leads to a special type of partial differential equation known as degenerate parabolic-elliptic or ultraparabolic [9].

For one-dimensional systems in the absence of external forces, the Hilbert space theory of the stationary transport equations with second-order differential collision operators has been recently developed to a quite satisfactory general level [1], [2]. Under certain regularity assumptions, the existence and uniqueness results for one-dimensional time-dependent problems follow immediately from the corresponding stationary theory (see [3], [7]). Since the results of [3] are based on (half-range) eigenfunction expansions, they appear to be limited to one-dimensional geometries, which allow for the separation of spatial and velocity variables in the streaming term of the transport equation.

The aim of this paper is to extend the existence and uniqueness theory for transport equations with second-order differential collision operators to a more general setting than considered previously. In § 2, we shall consider the existence of weak solutions, under boundary conditions of classical type, in L^P spaces, $1<p<\infty$. Time will not be singled out as in ordinary evolution problems, but rather treated together with position and velocity, which will allow inclusion of noncylindrical phase spaces and time-dependent transport operators.

In this paper, only bounded phase spaces have been considered, which leaves out the Fokker–Planck equation but includes the electron scattering equation. Existence theory in the spirit of Theorem 2.4 is probably feasible, even for unbounded domains, at the expense of extra technicalities. Much of the existence theory proceeds by duality arguments which single out L^1 as a distinct case. At this stage, it is not clear how to include it in the analysis.

Section 3 is devoted to the uniqueness issue. Since our definition of solution is rather weak, uniqueness is not always guaranteed. To recapture this feature, we have to strengthen the requirements on the solution. Some of these aspects of the uniqueness question are discussed in §§ 3 and 5. Section 4 extends the existence and uniqueness to include the case of general boundary conditions, encountered in transport problems.

In § 5, we construct a solution in a Hilbert space setting which provides uniqueness. A different Hilbert space setting also providing uniqueness is discussed in connection with recent results of Degond and Mas-Gallic [7] for the one-dimensional electron scattering problem.

2. Existence. In this section we shall discuss existence and a priori estimates for solutions of boundary value problems of classical type for transport equations with data in L^p. To be specific, we shall consider the operators

$$Lu \equiv \nabla_\xi \cdot \mu(\xi, x, t)\nabla_\xi u - \xi \cdot \nabla_x u + \xi \cdot \nabla_\xi u - \lambda u - \frac{\partial u}{\partial t} \tag{2.1}$$

and

$$L_0 u \equiv \nabla_\xi \cdot \mu_0(\xi, x)\nabla_\xi u - \xi \cdot \nabla_x u + \xi \cdot \nabla_\xi u - \lambda u \tag{2.2}$$

where $x \in \mathbb{R}^N$, $\xi \in \mathbb{R}^N$, $t \in \mathbb{R}$; ∇_ξ and ∇_x denote gradients taken with respect to ξ and x; μ and μ_0 are smooth, positive functions, and $\lambda > 0$ is a convenient positive constant. Due to the nature of the applications, we only consider first-order terms of the form shown in (2.1) and (2.2) (see the remarks at the end of this section for elaboration).

The problems we consider have the form $Lu = f$ in Ω with u specified on an appropriate subset of $\partial\Omega$ and similarly, $L_0 u = f$ on Ω_0 where Ω and Ω_0 are smooth, bounded domains in $\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}$ and $\mathbb{R}^N \times \mathbb{R}^N$, respectively. Observe that we can change λ in (2.1) by replacing u with $\tilde{u} = e^{-\alpha t} u$ for any constant α; since we shall study L_0 as the operator obtained by taking a given L and omitting the $\partial/\partial t$ term, we also have no essential loss of generality by taking a convenient choice of λ in L_0.

To describe the appropriate parts of the boundary of Ω or Ω_0 on which to specify boundary data, we shall use some ideas from the theory of second-order partial differential equations with nonnegative characteristic form. The ideas we use are modifications of those that were developed by G. Fichera and which are discussed in detail in the book of Oleinik and Radkevic [10]. Our modifications consist primarily of more detailed and, in some cases, more delicate analyses of the behavior of weak solutions on the boundary of domain. The extra information about the boundary values of our weak solutions will be used in § 4 to treat certain boundary conditions occurring from transport problems which are nontypical for second-order equations. Since our methods are the same for L and L_0, we give a detailed analysis only for L.

The notation used in [10] is somewhat different from that sometimes used in discussions of transport problems, such as [4]. The differences in notation reflect an interest focussed on different aspects of the operators L and L_0 and their physical interpretation. We shall give a brief discussion of how they are related and then state most of our results in the notation of [4], which emphasizes the *transport* character of L and L_0.

Suppose that $\Omega \subseteq \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}$ is a bounded domain with $\partial\Omega$ of class $C^{2,\alpha}$. We shall denote points of $\bar{\Omega}$ by (x, ξ, t). If $(x, \xi, t) \in \partial\Omega$, let $n(x, \xi, t) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}$ denote the *inner* normal to $\partial\Omega$ at (x, ξ, t). It will be convenient to write $n = (n_x, n_\xi, n_t)$ with $n_x \in \mathbb{R}^N$, $n_\xi \in \mathbb{R}^N$, and $n_t \in \mathbb{R}$. Following Fichera as presented in [10], we divide $\partial\Omega$ into four subsets. Let $\Sigma_3 = \{(x, \xi, t) \in \partial\Omega\colon n_\xi(x, \xi, t) \neq 0\}$. The set Σ_3 is the noncharacteristic part of $\partial\Omega$. Applying the definition given in [10] for general second-order operators to the specific operator L, let $b(x, \xi, t) = -\xi \cdot n_x + \xi \cdot n_\xi - n_t$ denote the Fichera function for L. Let

$$\Sigma_0 = \{(x, \xi, t) \in \partial\Omega \backslash \Sigma_3\colon b(x, \xi, t) = 0\},$$
$$\Sigma_1 = \{(x, \xi, t) \in \partial\Omega \backslash \Sigma_3\colon b(x, \xi, t) > 0\},$$
$$\Sigma_2 = \{(x, \xi, t) \in \partial\Omega \backslash \Sigma_3\colon b(x, \xi, t) < 0\}.$$

It can be shown that Σ_0, Σ_1, Σ_2, and Σ_3 are invariant under smooth nondegenerate changes of coordinates. If L^* denotes the formal adjoint of L, then the noncharacteristic part of $\partial\Omega$ for L^* is the same as for L; that is, $\Sigma_3(L^*) = \Sigma_3(L)$. On $\partial\Omega \backslash \Sigma_3$, applying the definition of [10] to L^* gives $b^* = -b$ where b^* is the Fichera function for L^*.

We now give the correspondence between some of the transport theory notation used in [4] and the notation of [10]. Let $D^+ = \Sigma_1$, $D^- = \Sigma_2$; let $d\nu^+$ denote the positive measure on D^+ given by $d\nu^+ = b\,d\sigma$ where $d\sigma$ the surface measure on $\partial\Omega$, and let $d\nu^- = -b\,d\sigma$ on D^-.

Our definition of weak solutions to boundary value problems for L and L_0 and our a priori estimates on these solutions are based on the following form of Green's identity: let $u, v \in C^2(\bar{\Omega})$; then

$$\int_\Omega (vLu - uL^*v)\,dx\,d\xi\,dt = -\int_{\Sigma_3} [vn_\xi \cdot \nabla_\xi u - un_\xi \cdot \nabla_\xi v]\mu\,d\sigma - \int_{\partial\Omega} buv\,d\sigma. \tag{2.3}$$

This is essentially formula (1.1.14) of [10]. If we require $u = v = 0$ on Σ_3 and use the notation of [4] we have the following lemma.

LEMMA 2.1. *Suppose* $u, v \in C^2(\bar{\Omega})$ *with* $u = v = 0$ *on* Σ_3. *Then*

$$\int_\Omega (vLu - uL^*v)\, dx\, d\xi\, dt = \int_{D^-} uv\, d\nu^- - \int_{D^+} uv\, d\nu^+. \tag{2.4}$$

Lemma 2.1 follows immediately from formula (1.1.14) of [10]. Formula (2.4) is closely related to formula (2.20) of [4]. (Note that in [4], $Y \simeq -L$ and formally $-Y^* = Y$; this accounts for the apparent sign difference between the two formulas.)

We can now state the basic problem considered in this section and give an appropriate weak formulation of the problem. We wish to solve

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Sigma_3, \\ u &= g \quad \text{on } D^- \end{aligned} \tag{2.5}$$

where we shall require $f \in L^p(\Omega)$, $g \in L^p(D^-, d\nu^-)$. Observe that if u is a classical solution of (2.5) then for any $v \in C^2(\bar{\Omega})$ with $v = 0$ on Σ_3 we have via (2.4) that

$$\int_\Omega (uL^*v)\, dx\, d\xi\, dt - \int_{D^+} uv\, d\nu^+ = \int_\Omega vf\, dx\, d\xi\, dt - \int_{D^-} vg\, d\nu^-. \tag{2.6}$$

To give a precise definition for the types of boundary data and weak solutions that we consider, we must first define certain spaces. Let

$$\begin{aligned} E^p &= L^p(\Omega, dx\, d\xi\, dt) \times L^p(D^+, d\nu^+) \times L^p(D^-, d\nu^-), \\ E_1^p &= L^p(\Omega, dx\, d\xi\, dt) \times L^p(D^+, d\nu^+) \times \{0\}, \\ E_2^p &= L^p(\Omega, dx\, d\xi\, dt) \times \{0\} \times L^p(D^-, d\nu^-), \end{aligned}$$

with norms

$$\|(u, u^+, u^-)\|_{E^p} = (\|u\|^p_{L^p(\Omega)} + \|u^+\|^p_{L^p(D^+, d\nu^+)} + \|u^-\|^p_{L^p(D^-, d\nu^-)})^{1/p}$$

and $\| \ \|_{E_1^p}$, $\| \ \|_{E_2^p}$ defined similarly. It is easy to see via the Riesz Representation Theorem (or its proof) that if $1 < p < \infty$ and $1/p + 1/q = 1$ then the dual spaces for E^p, E_1^p, and E_2^p are $E^{p*} = E^q$, $E_1^{p*} = E_1^q$, and $E_2^{p*} = E_2^q$.

Suppose $f \in L^p(\Omega)$ and $g \in L^p(D^-, d\nu^-)$.

DEFINITION. A weak solution of (2.5) in E^p is a triple $(u, u^+, u^-) \in E^p$ with $u^- = g$ such that

$$\int_\Omega uL^*v\, dx\, d\xi\, dt - \int_{D^+} u^+v^+\, d\nu^+ = \int_\Omega vf\, dx\, d\xi\, dt - \int_{D^-} v^-g\, d\nu^- \tag{2.7}$$

for all $v \in C^2(\bar{\Omega})$ with $v = 0$ on Σ_3, where $v^\pm = v|_{D^\pm}$.

Our objective in the remainder of this section will be to show that such weak solutions exist and satisfy certain a priori bounds. Specifically, it is crucial for the applications we consider to be able to control the L^p norm or u^+; our arguments are similar to those of [10] but are modified to allow us that control.

LEMMA 2.2. *Suppose that* $u \in C^2(\bar{\Omega})$ *with* $u = 0$ *on* Σ_3, *and* $1 < p < \infty$. *Then for* $\lambda > 0$

$$p\lambda \int_\Omega |u|^p\, dx\, d\xi\, dt + \int_{D^+} |u|^p\, d\nu^+ \leqq -p \int_\Omega |u|^{p-1} (\operatorname{sgn} u) Lu\, dx\, d\xi\, dt + \int_{D^-} |u|^p\, d\nu^- \tag{2.8}$$

and if we denote $u^+ = u|_{D^+}$ *and* $u^- = u|_{D^-}$, *then*

$$\lambda \|u\|^p_{L^p(\Omega)} + \|u^+\|^p_{L^p(D^+, d\nu^+)} \leqq \lambda^{1-p} \|Lu\|^p_{L^p(\Omega)} + \|u^-\|^p_{L^p(D^-, d\nu^-)}. \tag{2.9}$$

Proof. We follow the proof of Lemma 1.21 of [10], but do not require $u = 0^-$ on D^-. Applying (2.3) to the pair of functions $\{(u^2+\delta)^{p/2}, -1\}$ yields

$$\begin{aligned}\int_\Omega \lambda(u^2+\delta)^{p/2}\,dx\,d\xi\,dt + \int_\Omega L(u^2+\delta)^{p/2}\,dx\,d\xi\,dt \\ = -\int_{\Sigma_3} b(u^2+\delta)^{p/2}\,d\sigma + \int_{D^-}(u^2+\delta)^{p/2}\,d\nu^- - \int_{D^+}(u^2+\delta)^{p/2}\,d\nu^+.\end{aligned} \tag{2.10}$$

(Observe that since $u = 0$ on Σ_3, we have

$$\nabla_\xi (u^2+\delta)^{p/2} = (p/2)(u^2+\delta)^{(p/2-1)}(2u\nabla_\xi u) = 0 \quad \text{on } \Sigma_3;$$

since $v = -1$ is constant the first integral on the right side of (2.3) is zero.) A calculation yields

$$\begin{aligned}L(u^2+\delta)^{p/2} = p(u^2+\delta)^{p/2-1}uLu - \lambda(u^2+\delta)^{p/2-1}[(1-p)u^2+\delta] \\ + \mu(\xi, x, t)(\nabla_\xi u \cdot \nabla_\xi u)p(u^2+\delta)^{p/2-2}[(p-1)u^2+\delta];\end{aligned}$$

since the last term on the right is nonnegative and since $u = 0$ on Σ_3, we obtain from (2.10) the inequality

$$\begin{aligned}\int_\Omega \{\lambda(u^2+\delta)^{p/2} - \lambda(u^2+\delta)^{p/2-1}[(1-p)u^2+\delta]\}\,dx\,d\xi\,dt \\ + \int_\Omega p(u^2+\delta)^{p/2-1}uLu\,dx\,d\xi\,dt \leqq \delta^{p/2}\int_{\Sigma_3}|b|\,d\sigma + \int_{D^-}(u^2+\delta)^{p/2}\,d\nu^- \\ - \int_{D^+}(u^2+\delta)^{p/2}\,d\nu^+.\end{aligned}$$

Letting $\delta \to 0$ yields

$$p\lambda\int_\Omega |u|^p\,dx\,d\xi\,dt + p\int_\Omega |u|^{p-2}\,uLu\,dx\,d\xi\,dt \leqq \int_{D^-}|u|^p\,d\nu^- - \int_{D^+}|u|^p\,d\nu^+,$$

which is equivalent to (2.8). To obtain (2.9) from (2.8) we use Young's inequality, as in the proof of Proposition 4 in [4, § 3]: if $a, b > 0$ and $1 < p < \infty$, $1/p + 1/q = 1$, we have $ab \leqq a^p/p + b^q/q$. Applying the inequality with $a = \lambda^{-1/q}|Lu|$, $b = \lambda^{1/q}|u|^{p-1}$ and integrating over Ω yields

$$\begin{aligned}p\int_\Omega |u|^{p-1}|Lu|\,dx\,d\xi\,dt \leqq p\int_\Omega (\lambda^{-p/q}|Lu|^p/p)\,dx\,d\xi\,dt \\ + p\int_\Omega (\lambda^{q/q}|u|^{(p-1)q}/q)\,dx\,d\xi\,dt \\ = \lambda^{1-p}\|Lu\|^p_{L^p(\Omega)} + \lambda(p-1)\|u\|^p_{L^p(\Omega)}.\end{aligned}$$

Using the last inequality to estimate the first term on the right side in (2.8) and combining like terms yields (2.9). □

The same type of analysis may be applied to L^*. Noting that the sign of b changes and thus the role of D^+ and D^- are reversed for L^*, we have Lemma 2.3.

LEMMA 2.3. *Suppose that* $v \in C^2(\bar{\Omega})$ *with* $v = 0$ *on* Σ_3, *and* $1 < q < \infty$. *Then for* $\lambda > 0$,

$$q\lambda \int_\Omega |v|^q \, dx \, d\xi \, dt + \int_{D^-} |v|^q \, d\nu^- \leqq -q \int_\Omega |v|^{q-1} (\operatorname{sgn} v) L^* v \, dx \, d\xi \, dt + \int_{D^+} |v|^q \, d\nu^+ \tag{2.11}$$

and if $v^+ = v|_{D^+}$, $v^- = v|_{D^-}$, *then*

$$\lambda \|v\|^q_{L^q(\Omega)} + \|v^-\|^q_{L^q(D^-, d\nu^-)} \leqq \lambda^{1-q} \|L^* v\|^q_{L^q(\Omega)} + \|v^+\|^q_{L^q(D^+, d\nu^+)}. \tag{2.12}$$

We now turn to the question of existence of solutions to (2.5) in the sense of (2.7).

THEOREM 2.4. *Suppose that* $\lambda \geqq 1$, $1 < p < \infty$, $f \in L^p(\Omega)$, *and* $g \in L^p(D^-, d\nu^-)$. *Then* (2.5) *has a solution* $(u, u^+, u^-) \in E^p$ *in the sense of* (2.7) *with* $u^- = g$, *and that solution satisfies the estimate*

$$\inf_{(y,z,0) \in Z} [\|u + y\|^p_{L^p(\Omega)} + \|u^+ - z\|^p_{L^p(D^+, d\nu^+)}] \leqq \|f\|^p_{L^p(\Omega)} + \|g\|^p_{L^p(D^-, d\nu^-)} \tag{2.13}$$

where

$$Z = \left\{ (y, z, 0) \in E_1^p : \int_\Omega y L^* v \, dx \, d\xi \, dt + \int_{D^+} z v^+ \, d\nu^+ = 0 \right.$$
$$\left. \text{for all } v \in C^2(\bar{\Omega}) \text{ with } v = 0 \text{ on } \Sigma_3 \text{ and } v^+ = v|_{D^+} \right\}.$$

Remark. If $Z = \{0\}$, then (2.13) becomes

$$\|u\|^p_{L^p(\Omega)} + \|u^+\|^p_{L^p(D^+, d\nu^+)} \leqq \|f\|^p_{L^p(\Omega)} + \|g\|^p_{L^p(D^-, d\nu^-)}. \tag{2.14}$$

Conditions implying $Z = \{0\}$ are given in the next section. The dimension or "size" of the subspace Z measures the nonuniqueness of solutions in our sense in a given L^p class. That Z need not always be $\{0\}$ is illustrated in the next section via a counter example adapted from [10].

Proof. Let q be such that $1/p + 1/q = 1$. Since $\lambda \geqq 1$ it follows from (2.12) of Lemma 2.3 that for $v \in C^2(\bar{\Omega})$ with $v = 0$ on Σ_3 we have

$$\|(v, 0, v^-)\|^q_{E^q} \leqq \|(L^* v, v^+, 0)\|^q_{E^q} \quad \text{where } v^\pm = v|_{D^\pm}. \tag{2.15}$$

Let $\tilde{E}_1^q$ be the completion in E_1^q of the subspace $\{(L^* v, v^+, 0) : v \in C^2(\bar{\Omega}), v = 0 \text{ on } \Sigma_3\}$. If $(\varphi, \psi_1, 0) \in \tilde{E}_1^q$ then there must exist a sequence $v_n \in C^2(\bar{\Omega})$ with $v_n = 0$ on Σ_3 such that $(L^* v_n, v_n^+, 0) \to (\varphi, \psi_1, 0)$ in E_1^q (where $v_n^+ = v_n|_{D^+}$), so the sequence $\{(v_n, 0, v_n^-)\}$ with $v_n^- = v_n|_{D^-}$ is Cauchy in E_2^q by (2.15), and we obtain a uniquely defined limit $(\theta, 0, \psi_2) \in E_2^q$ satisfying

$$\|(\theta, 0, \psi_2)\|_{E_2^q} \leqq \|(\varphi, \psi_1, 0)\|_{E_1^q}. \tag{2.16}$$

Thus, the map $(L^* v, v^+, 0) \mapsto (v, 0, v^-)$ extends to a bounded linear map $G : \tilde{E}_1^q \to E_2^q$ with $\|G\| \leqq 1$. Suppose $(\varphi, \psi_1, 0) \in \tilde{E}_1^q$ with $G(\varphi, \psi_1, 0) = (\theta, 0, \psi_2)$. Define a linear functional F on $\tilde{E}_1^q$ as follows:

$$F(\varphi, \psi_1, 0) = \int_\Omega \theta f \, dx \, d\xi \, dt - \int_{D^-} \psi_2 g \, d\nu^-. \tag{2.17}$$

By Hölder's inequality, we have

$$\begin{aligned}|F(\varphi,\psi_1,0)| &\leqq \|\theta\|_{L^q(\Omega)}\|f\|_{L^p(\Omega)}+\|\psi_2\|_{L^q(D^-,d\nu^-)}\|g\|_{L^p(D^-,d\nu^-)}\\ &\leqq [\|\theta\|^q_{L^q(\Omega)}+\|\psi_2\|^q_{L^q(D^-,d\nu^-)}]^{1/q}[\|f\|^p_{L^p(\Omega)}+\|g\|^p_{L^p(D^-,d\nu^-)}]^{1/p}\\ &= \|(\theta,0,\psi_2)\|_{E_2^q}\|(f,0,g)\|_{E_2^p}\\ &= \|G(\varphi,\psi_1,0)\|_{E_2^q}\|(f,0,g)\|_{E_2^p}\\ &\leqq \|(\varphi,\psi_1,0)\|_{E_1^q}\|(f,0,g)\|_{E_2^p}.\end{aligned}$$

Thus, F is a bounded linear functional on $\tilde{E}_1^q$ with $\|F\|\leqq\|(f,0,g)\|_{E_2^p}$. Since $\tilde{E}_1^q$ is a closed subspace of E_1^q, it follows from the Hahn–Banach theorem that we may extend F to a functional $\hat{F}$ on all of E_1^q with $\|\hat{F}\|\leqq\|(f,0,g)\|_{E_2^p}$. In general, the extension $\hat{F}$ of F and the element of $E_1^p\cong E_1^{q*}$ associated with $\hat{F}$ are not unique. In fact, $\tilde{E}_1^{q*}$ can be identified with the factor space E_1^p/Z where

$$Z=\left\{(y,z,0)\in E_1^p\colon \int_\Omega y\varphi\,dx\,d\xi\,dt+\int_{D^+} z\psi_1 d\nu^+=0 \text{ for all } (\varphi,\psi_1,0)\in\tilde{E}_1^q\right\}.$$

Since $\tilde{E}_1^q$ is the closure of $\{(L^*v,v^+,0)\colon v\in C^2(\bar{\Omega}),\ v=0 \text{ on } \Sigma_3\}$, we may characterize Z as

$$Z=\left\{(y,z,0)\in E_1^p\colon \int_\Omega yL^*v\,dx\,d\xi\,dt+\int_{D^+} zv^+\,d\nu^+=0 \quad \text{for all } v\in C^2(\bar{\Omega}) \text{ with } v=0 \text{ on } \Sigma_3\right\}.$$

From the identification of $\tilde{E}_1^{q*}$ with E_1^p/Z it follows that F can be represented by a coset $[(u,w,0)+Z]\in E_1^p/Z$; that is, for $(\varphi,\psi_1,0)\in\tilde{E}_1^q$ and $(\theta,0,\psi_2)=G(\varphi,\psi_1,0)$ we have

$$\int_\Omega \theta f\,dx\,d\xi\,dt-\int_{D^-}\psi_2 g\,d\nu^-=F(\varphi,\psi_1,0)=\int_\Omega u\varphi\,dx\,d\xi\,dt+\int_{D^+} w\psi_1\,d\nu^+. \tag{2.18}$$

If $(\varphi,\psi_1,0)=(L^*v,v^+,0)$ for some $v\in C^2(\bar{\Omega})$ with $v=0$ on Σ_3, (2.18) becomes (taking $v^\pm=v|_{D^\pm}$)

$$\int_\Omega vf\,dx\,d\xi\,dt-\int_{D^-} v^-g\,d\nu^-=\int_\Omega uL^*v\,dx\,d\xi\,dt+\int_{D^+} wv^+\,d\nu^+. \tag{2.19}$$

If we let $u^+=-w$ and $u^-=g$, then (2.19) is equivalent to (2.7). Also, $\|F\|_{\tilde{E}_1^{q*}}\leqq\|(f,0,g)\|_{E_2^p}$, and since we may identify E_1^{q*} with the factor space E_1^p/Z it follows from the definition of the norm for a coset that

$$\inf_{(y,z,0)\in Z}\|(u,w,0)+(y,z,0)\|_{E_1^p}=\|F\|\leqq\|(f,0,g)\|_{E_2^p}. \tag{2.20}$$

By the definition of the norms on E_1^p and E_2^p, and using $u^+=-w$, (2.20) is equivalent to (2.13). □

The analysis for L_0 is essentially the same as that for L, so we shall only state the corresponding results. Suppose that $\Omega_0\subseteq\mathbb{R}^N\times\mathbb{R}^N$ is bounded with $\partial\Omega_0$ of class $C^{2,\alpha}$. Let $n(x,\xi)=(n_x,n_\xi)\in\mathbb{R}^N\times\mathbb{R}^N$ denote the inner unit normal at (x,ξ). Let

$$\Sigma_3^0=\{(x,\xi)\in\partial\Omega\colon n_\xi(x,\xi)\neq 0\} \quad \text{let } b_0=-\xi\cdot n_x+\xi\cdot n_\xi,$$

$$D_0^+=\{(x,\xi)\in\partial\Omega\backslash\Sigma_3^0\colon b_0(x,\xi)>0\},$$

$$D_0^-=\{(x,\xi)\in\partial\Omega\backslash\Sigma_3^0;\ b_0(x,\xi)<0\} \quad \text{and} \quad d\nu_0^\pm=\pm b_0\,d\sigma \quad \text{on } D_0^\pm.$$

Define the spaces

$$E_0^p = L^p(\Omega_0) \times L^p(D_0^+, d\nu_0^+) \times L^p(D_0^-, d\nu_0^-),$$
$$E_{01}^p = L^p(\Omega_0) \times L^p(D_0^+, d\nu_0^+) \times \{0\},$$
$$E_{02}^p = L^p(\Omega_0) \times \{0\} \times L^p(D_0^-, d\nu_0^-),$$

with norms as for E^p, E_1^p, and E_2^p. We are interested in solutions in E_0^p of

$$\begin{aligned} L_0 u &= f \quad \text{in } \Omega_0, \\ u &= 0 \quad \text{on } \Sigma_3^0, \\ u &= g \quad \text{on } D_0^- \end{aligned} \tag{2.21}$$

with $(f, 0, g) \in E_{02}^p$. Such solutions are defined as for (2.5); that is, $(u, u^+, u^-) \in E_0^p$ is a solution of (2.21) if $u^- = g$ and for every $v \in C^2(\bar{\Omega}_0)$ with $v = 0$ on Σ_3^0 we have

$$\int_{\Omega_0} u L_0^* v \, dx \, d\xi - \int_{D_0^+} u^+ v^+ \, d\nu_0^+ = \int_{\Omega_0} v f \, dx \, d\xi - \int_{D_0^-} v^- g \, d\nu_0^- \tag{2.22}$$

where $v^\pm = v|_{D_0^\pm}$.

Corresponding to our results for (2.5) we have the following lemma.

LEMMA 2.5. *Suppose that $u \in C^2(\bar{\Omega}_0)$ with $u = 0$ on Σ_3^0 and $1 < p < \infty$. Then for $\lambda > 0$,*

$$p\lambda \int_{\Omega_0} |u|^p \, dx \, d\xi + \int_{D_0^+} |u|^p \, d\nu_0^+ \leqq -p \int_{\Omega_0} |u|^{p-1} \operatorname{sgn}(u) L_0 u \, dx \, d\xi + \int_{D_0^-} |u|^p \, d\nu_0^-, \tag{2.23}$$

$$p\lambda \int_{\Omega_0} |u|^p \, dx \, d\xi + \int_{D_0^-} |u|^p \, d\nu_0^- \leqq -p \int_{\Omega_0} |u|^{p-1} \operatorname{sgn}(u) L_0^* u \, dx \, d\xi + \int_{D_0^+} |u|^p \, d\nu_0^+. \tag{2.24}$$

Moreover if we denote $u^\pm = u|_{D_0^\pm}$, then

$$\lambda \|u\|_{L^p(\Omega_0)}^p + \|u^+\|_{L^p(D_0^+, d\nu_0^+)}^p \leqq \lambda^{1-p} \|L_0 u\|_{L^p(\Omega_0)}^p + \|u^-\|_{L^p(D_0^-, d\nu_0^-)}^p \tag{2.25}$$

and

$$\lambda \|u\|_{L^p(\Omega_0)}^p + \|u^-\|_{L^p(D_0^-, d\nu_0^-)}^p \leqq \lambda^{1-p} \|L_0^* u\|_{L^p(\Omega_0)}^p + \|u^+\|_{L^p(D_0^+, d\nu_0^+)}^p. \tag{2.26}$$

Using the estimates of Lemma 2.5 yields the following existence result via essentially the same proof as that of Theorem 2.4.

THEOREM 2.6. *Suppose that $\lambda \geqq 1$, $1 < p < \infty$, $f \in L^p(\Omega_0)$, and $g \in L^p(D_0^-, d\nu_0^-)$. Then (2.21) has a solution $(u, u^+, u^-) \in E_0^p$ in the sense of (2.22) with $u^- = g$ and that solution satisfies the estimate*

$$\inf_{(y,z,0) \in Z_0} [\|u + y\|_{L^p(\Omega_0)}^p + \|u^+ - z\|_{L^p(D_0^+, d\nu_0^+)}^p] \leqq \|f\|_{L^p(\Omega_0)}^p + \|g\|_{L^p(D_0^-, d\nu_0^-)}^p \tag{2.27}$$

where

$$Z_0 = \left\{ (y, z, 0) \in E_{01}^p : \int_{\Omega_0} y L^* v \, dx \, d\xi + \int_{D_0^+} z v^+ \, d\nu^+ = 0 \right.$$
$$\left. \text{for all } v \in C^2(\bar{\Omega}) \text{ with } v = 0 \text{ on } \Sigma_3^0 \text{ and } v^+ = v|_{D_0^+} \right\}.$$

Remark. Again, if $Z_0 = \{0\}$, (2.27) becomes

$$\|u\|_{L^p(\Omega_0)}^p + \|u^+\|_{L^p(D_0^+, d\nu_0^+)}^p \leqq \|f\|_{L^p(\Omega_0)}^p + \|g\|_{L^p(D_0^-, d\nu_0^-)}^p. \tag{2.28}$$

As noted in the remarks following Theorem 2.4, Z_0 provides a measure of the nonuniqueness of our weak solutions.

In general, estimates of the type given in (2.14) or (2.28) are not quite strong enough to permit a free application of semigroup theory. A more careful analysis can provide somewhat sharper estimates in terms of λ when $g=0$, but those are not quite sufficient. To obtain a weak solution satisfying strong enough estimates we must return to the original formulation of weak solutions on L^p given in [10], and thereby relinquish some control over the boundary behavior of our solutions.

In the case $g=0$, if we return to the analysis following (2.17) we have

$$|F(\varphi,\psi_1,0)|\leqq\|\theta\|_{L^q(\Omega)}\|f\|_{L^p(\Omega)}\leqq[\lambda\|\theta\|^q_{L^q(\Omega)}]^{1/q}\lambda^{-1/q}\|f\|_{L^p(\Omega)}. \tag{2.29}$$

Since there exist $v_n\in C^2(\bar{\Omega})$ with $v_n=0$ on Σ_3 such that $(L^*v_n, v_n^+, 0)\to(\varphi,\psi_1,0)$ in E_1^q and $(v_n, 0, v_n^-)\to(\theta,0,\psi_2)$ in E_2^q, and since each v_n satisfies (2.12), we have

$$\lambda\|\theta\|^q_{L^q(\Omega)}+\|\psi_2\|^q_{L^q(D^-,d\nu^-)}\leqq\lambda^{1-q}\|\varphi\|^q_{L^q(\Omega)}+\|\psi_1\|^q_{L^q(D^+,d\nu^+)}$$

so that for $\lambda\geqq1$ (2.29) yields

$$\begin{aligned}|F(\varphi,\psi_1,0)|&\leqq[\lambda^{1-q}\|\varphi\|^q_{L^q(\Omega)}+\|\psi_1\|^q_{L^q(D^+,d\nu^+)}]^{1/q}\lambda^{-1/q}\|f\|_{L^p(\Omega)}\\&\leqq[\|\varphi\|^q_{L^q(\Omega)}+\|\psi_1\|^q_{L^q(D^+,d\nu^+)}]^{1/q}\lambda^{-1/q}\|f\|_{L^p(\Omega)}\end{aligned}$$

so that

$$|F(\varphi,\psi_1,0)|\leqq\|(\varphi,\psi_1,0)\|_{E_1^q}\lambda^{-1/q}\|f\|_{L^p(\Omega)}$$

and thus

$$\|F\|\leqq\lambda^{-1/q}\|f\|_{L^p(\Omega)}. \tag{2.30}$$

Using the bound (2.30) in the remainder of the proof of Theorem 2.4 allows us to replace (2.14) with the estimate

$$\|u\|^p_{L^p(\Omega)}+\|u^+\|^p_{L^p(D^+,d\nu^+)}\leqq\lambda^{-p/q}\|f\|^p_{L^p(\Omega)} \tag{2.31}$$

provided $\lambda\geqq1$, $g=0$, and $Z=\{0\}$. Similarly, if $\lambda\geqq1$, $Z_0=\{0\}$ and $g=0$ in Theorem 2.6, we may replace (2.28) with

$$\|u\|^p_{L^p(\Omega_0)}+\|u^+\|^p_{L^p(D_0^+,d\nu_0^+)}\leqq\lambda^{-p/q}\|f\|^p_{L^p(\Omega_0)}. \tag{2.32}$$

To apply semigroup theory to L_0 we would really want an estimate on $\|(u,u^+,0)\|_{E_{01}^p}$ of order $\lambda^{-1}\|f\|_{L^p(\Omega_0)}$. We can obtain an estimate on $\|u\|_{L^p(\Omega_0)}$ for certain weak solutions from the theory of [10], but only at the price of losing essentially all information about u^+, which largely defeats the purpose of the present article. Specifically, in [10] the L^p weak solutions for

$$\begin{aligned}&L_0u=f \quad \text{in } \Omega_0,\\&u=0 \quad \text{on } \Sigma_3^0,\\&u=0 \quad \text{on } D_0^+\end{aligned} \tag{2.33}$$

are required to satisfy

$$\int_{\Omega_0}uL_0^*v\,dx\,d\xi=\int_{\Omega_0}vf\,dx\,d\xi \tag{2.34}$$

for all $v\in C^2(\bar{\Omega})$ with $v=0$ on $\Sigma_3^0\cup D_0^+$. The following is essentially Theorem 1.3.1 of [10] specialized to L_0.

THEOREM 2.7. *Suppose that* $\lambda \geqq 1$, $1<p<\infty$, *and* $f \in L^p(\Omega_0)$. *Then* (2.33) *has a solution* $u \in L^p(\Omega_0)$ *in the sense of* (2.34) *and the solution satisfies the estimate*

$$\inf_{y \in \tilde{Z}_0} \|u+y\|_{L^p(\Omega_0)} \leqq \lambda^{-1} \|f\|_{L^p(\Omega_0)} \tag{2.35}$$

where

$$\tilde{Z}_0 = \left\{ y \in L^p(\Omega_0) : \int_{\Omega_0} y L_0^* v \, dx \, d\xi = 0 \text{ for all } v \in C^2(\bar{\Omega}) \text{ with } v = 0 \text{ on } \Sigma_0^3 \cup D_0^+ \right\}.$$

Remarks. In the situation where $\tilde{Z}_0 = \{0\}$ we have, from (2.35),

$$\|u\|_{L^p(\Omega_0)} \leqq \lambda^{-1} \|f\|_{L^p(\Omega_0)}. \tag{2.36}$$

Other types of possible weak solutions and conditions under which these solutions coincide are discussed at the end of the next section. Note however that since the test functions v in (2.34) are required to be zero on D_0^+, that formulation of a weak solution gives no information about the existence or properties of a trace u^+ for u on D_0^+.

We could replace the coefficients of the first-order terms in (2.1) by general (sufficiently smooth) functions. This replacement is necessary if one wishes to include the effects of external forces, via a term of the form $\nu(x, \xi, t) \nabla_\xi u$. The replacement of the other coefficients of the first-order terms is possible, but academic, as far as *transport* problems are concerned, since their significance is related to the vector field part of the transport equation. Either replacement changes only the actual form of the characteristic parts of the boundary and the corresponding results follow similarly.

3. Uniqueness. We now consider the question of uniqueness for the solutions obtained in Theorems 2.4 and 2.6. Clearly, the solutions obtained are unique if and only if $Z = \{0\}$ and $Z_0 = \{0\}$; and in that case the estimates (2.14) and (2.28) hold. To conclude uniqueness we must impose certain additional conditions on f, g, and Ω or Ω_0; that some type of additional hypotheses are needed is illustrated by a counterexample at the end of this section.

Suppose that $\partial\Omega$ is given locally by the equation $h(x, \xi, t) = 0$ with $(\nabla_x h, \nabla_\xi h, \partial h/\partial t) \neq 0$ and $h > 0$ inside Ω. Define $\beta^* \equiv L^* h$. At points of $\partial\Omega$ which lie in the interior of $\partial\Omega \backslash \Sigma_3$ (or at limits of sequences of such points) the sign of β^* agrees with that of $b^* = -b$, the Fichera function for L^* (see [10, p. 31]). Let Γ denote the boundary in $\partial\Omega$ of the set $D^- \cup \Sigma_0$. We have the following uniqueness result, which is a special case of Theorem 1.6.1 of [10].

LEMMA 3.1. *Suppose that* $\lambda > 0$ *and that* $\beta^* < 0$ *at points of* D^+. *Suppose that in some neighborhood of each of its points* Γ *lies on the intersection of the surface* $h(x, \xi, t) = 0$ *defining* $\partial\Omega$ *and a surface* $\Psi(x, \xi, t) = 0$ *with* h *and* Ψ *of class* C^2 *such that the normal* h *to* $\partial\Omega$ *is not orthogonal to the surface* $\Psi(x, \xi, t) = 0$. *If* $u \in L^p(\Omega)$ *for* $p \geqq 3$ *and*

$$\int_\Omega u L^* v \, dx \, d\xi \, dt = 0 \tag{3.1}$$

for every $v \in C^2(\bar{\Omega})$ *with* $v = 0$ *on* $\Sigma_3 \cup D^+$, *then* $u = 0$ *almost everywhere in* Ω.

Remark. If $\partial\Omega \backslash \Sigma_3$ consists entirely of its own interior points in $\partial\Omega$ and their limits, then as noted before the statement of Lemma 3.1 the condition $\beta^* < 0$ on D^+ is satisfied automatically since $b < 0$ on D^+.

Lemma 3.1 yields the following uniqueness result.

THEOREM 3.2. *Suppose that* Ω *satisfies the conditions of Lemma* 3.1, $\lambda \geqq 1$, *and* $3 \leqq p < \infty$. *Then if* Z *is the set defined in Theorem* 2.4, *we have* $Z = \{0\}$, *so* (2.5) *has a unique weak solution* $(u, u^+, u^-) \in E^p$ *and the weak solution satisfies* (2.14).

Proof. Suppose $(y, z, 0) \in Z$. Then by definition

$$\int_\Omega yL^*v\, dx\, d\xi\, dt + \int_{D^+} zv^+\, d\nu^+ = 0$$

for any $v \in C^2(\bar{\Omega})$ with $v = 0$ on Σ_3. If $v = 0$ on D^+ then $v^+ = 0$ so for $v \in C^2(\bar{\Omega})$ with $v = 0$ on $\Sigma_3 \cup D^+$ we have

$$\int_\Omega yL^*v\, dx\, d\xi\, dt = 0.$$

Since $(y, z, 0) \in Z \subseteq E^p$, $p \geqq 3$ and Ω satisfies the hypotheses of Lemma 3.1, it follows from that lemma that $y = 0$ almost everywhere in Ω. Hence, we have

$$\int_{D^+} v^+ z\, d\nu^+ = 0$$

for all $v^+ = v|_{D^+}$ with $v \in C^2(\bar{\Omega})$ and $v = 0$ on Σ_3. Since $C_0^\infty(D^+)$ is dense in $L^q(D^+, d\nu^+)$ for any q, we have $z = 0$ almost everywhere on D^+. Thus we see that $Z = \{0\}$. If (u_1, u_1^+, u_1^-) and (u_2, u_2^+, u_2^-) are solutions of (2.5) in E^p then $u_1^- = u_2^- = g$ and $(u_1 - u_2, u_1^+ - u_2^+, 0) \in Z$, so $Z = \{0\}$ implies $u_1 = u_2$ almost everywhere in Ω and $u_1^+ = u_2^+$ almost everywhere $(d\nu^+)$ in D^+. Hence $(u_1, u_1^+, u_1^-) = (u_2, u_2^+, u_2^-)$ in E^p, establishing the uniqueness of the weak solution. Finally, (2.14) immediately follows from (2.13) once we know that $Z = \{0\}$. □

If the boundary of Ω_0 is given locally by $h_0(x, \xi) = 0$, then we can define $\beta_0^* = L_0^* h_0$, and define Γ_0 to be the boundary in $\partial\Omega_0$ of $D_0^+ \cup \{(x, \xi) \in \partial\Omega_0 \backslash \Sigma_3^0 \colon b_0(x, \xi) = 0\}$. Applying the results of [10] to L_0 in the same way we applied them to L, we obtain Theorem 3.3.

THEOREM 3.3. *Suppose that $\lambda \geqq 1$ and $\beta_0^* < 0$ at points of D_0^+. Suppose that in some neighborhood of each of its points Γ_0 lies on the intersection of the surface $h_0(x, \xi) = 0$ defining $\partial\Omega_0$ and a surface $\Psi_0(x, \xi) = 0$ with h_0 and Ψ of class C^2 such that the normal to $\partial\Omega$ is not orthogonal to $\Psi(x, \xi) = 0$. Then for $3 \leqq p < \infty$ we have $Z_0 = \{0\}$, where Z_0 is the set defined in Theorem* 2.6, *and hence problem* (2.21) *has a unique weak solution $(u, u^+, u^-) \in E_0^p$, and that solution satisfies* (2.28).

We now present an example showing that the condition $p \geqq 3$ in Theorem 3.3 as well as in Theorem 3.2 is sharp. In other words the solution may not be unique in E^p. As a particular realization of (2.1), we consider the equation

$$u_{yy} + yu_y - u_s = 0$$

in a bounded domain in the (y, s) plane with $s > 0$. By making the change of coordinates $y = \xi/\sqrt{2t}$, $s = (\ln t)/2$ (equivalently $t = e^{2s}$, $\xi = \sqrt{2}e^s y$) we obtain the heat equation

$$u_{\xi\xi} - u_t = 0. \tag{3.2}$$

Following [10], we consider this last equation in a bounded domain Ω in the plane (ξ, t) such that the boundary of Ω is a closed smooth curve Σ which in the neighborhood of $(0, 0)$ behaves like $t = |\xi|^{2+\varepsilon}$, $\varepsilon > 0$, and which has no tangents parallel to the ξ-axis except at $(0, 0)$ and $(0, 1)$. Then the points $(0, 0)$, $(0, 1)$ belong to D^- and D^+, respectively, and the rest of the boundary belongs to Σ_3. The function $w = t^{-1/2} e^{-\xi^2/4t}$ is a solution of (3.2) in Ω. On Σ, w determines a continuous function which in a neighborhood of $(0, 0)$ behaves like

$$|\xi|^{-(1+\varepsilon/2)} e^{-1/4|\xi|^{-\varepsilon}}.$$

We now consider the problem

$$u_{\xi\xi} = u_t = 0, \qquad u|_\Sigma = w|_\Sigma. \tag{3.3}$$

Since the boundary Σ is smooth, there exists a solution $W(\xi, t)$ of problem (3.3), continuous together with its derivatives W_t, W_ξ, $W_{\xi\xi}$, everywhere in $\Omega \cup \Sigma$ except maybe at $(0,0)$ and $(0,1)$ [8, Thms. 5, 6, p. 64].

The function $u := W - w$ satisfies the homogeneous equation

$$L(u) = 0, \qquad u|_\Sigma = 0, \tag{3.4}$$

but is different from zero on a set of positive measure, since W is continuous on $\Omega \cup \Sigma$ while $w \to \infty$ for $\xi = 0$ and $t \to 0$. Moreover, $u \in L^p(\Omega)$ if $p < 3$. Indeed, since W is bounded, we have

$$\begin{aligned}\int_\Omega |u|^p \, d\xi \, dt &\simeq \int_\Omega |w|^p \, d\xi \, dt \sim \int t^{-p/2} \, e^{-p\xi^2/4t} \, d\xi \, dt \\ &\leqq \int_0^{t_0} \left(t^{-p/2} \int_{-t^{1/(2+\varepsilon)}}^{t^{1/(2+\varepsilon)}} d\xi \right) dt + C_1 \\ &= 2 \int_0^{t_0} t^{-p/2+1/(2+\varepsilon)} \, dt + C_1 < C_2\end{aligned}$$

if $-p/2 + 1/(2+\varepsilon) > -1$, i.e., if $p < 3$.

Following [10] it is not difficult to check that u also satisfies Green's identity; therefore, it is a solution of (3.4) different from the one that is identically zero.

The lack of uniqueness comes from our rather general definition of a solution. Extra regularity requirements can restore this property. For instance, when $p \geqq 3$ our formulation already imposes enough regularity to ensure uniqueness. A different formulation that imposes more regularity is given in § 5, and yields a uniqueness result for data in L^2. When we have uniqueness we may assert that weak solutions obtained by different methods must coincide. That observation is useful because certain properties of the solution may be easier to obtain by some methods than by others.

Our approach was chosen to yield information about u^+ and u^- and their relation to u and f. Another approach, used in [10], can be adapted to yield more precise information about the sign of solutions. The alternative method is to approximate f and g with sequences with smooth functions f_n and g_n such that $f_n \to f$ in $L^p(\Omega)$, $g_n \to g$ in $L^p(D^-, d\nu^-)$, and then to solve the problem

$$\begin{aligned} &\varepsilon \Delta u + Lu = f_n && \text{in } \Omega, \\ &u = g_n && \text{on } D^-, \\ &u = 0 && \text{on } \partial\Omega \setminus D^- \end{aligned} \tag{3.5}$$

where $\Delta u = \nabla_\xi \cdot \nabla_\xi u + \nabla_x \cdot \nabla_x u + u_{tt}$. Standard elliptic theory asserts the evidence of a classical solution to (3.5) for each ε and n, and it can be shown as in [10, § 1.5], that if the solution to (3.5) is denoted by $u_{\varepsilon,n}$ then there are subsequences $\varepsilon_k \to 0$ and $n_k \to \infty$ such that u_{ξ_k, n_k} converges weakly to $\tilde{u} \in L^p(\Omega)$, with $\tilde{u}$ satisfying

$$\int_\Omega \tilde{u} L^* v \, dx \, d\xi \, dt = \int_\Omega v f \, dx \, d\xi \, dt - \int_{D^-} v g \, d\nu^- \tag{3.6}$$

for all $v \in C^2(\bar{\Omega})$ with $v = 0$ on $\Sigma_3 \cup D^+$. (It seems to be difficult to adapt this approach to the formulation given in (2.7) where v need not vanish on D^+. Since we are interested in transport phenomena where the relation between the characteristic boundary values u^- and u^+ is essential, our main line of analysis uses the alternative approach discussed in § 2.) The advantage of the above method is that if $f \leqq 0$ and $g \geqq 0$ almost everywhere,

then we can construct f_n and g_n by mollification so that $f_n \leqq 0$ and $g_n \geqq 0$ for each ε and n, so that $\tilde{u} \geqq 0$ almost everywhere in Ω. Under the hypotheses of Theorem 3.2, the weak solution $\tilde{u}$ in the sense of (3.6) is unique. However, our weak solution u in the sense of (2.7) is also unique, and satisfies (3.6). Hence, under the hypotheses of Theorem 3.2, $u = \tilde{u} \geqq 0$ almost everywhere.

Another approach to solving (2.5) in a domain of the form $\Omega_0 \times (0, T]$ with $g = 0$ on $D_0^- \times (0, T]$ would be to assume $\mu = \mu(x, \xi)$, use the estimate (2.36) and the Hille-Yosida Theorem to assert that for $p \geqq 3$, L_0 generates a C_0-semigroup of contractions on $L^p(\Omega_0)$, and then to use the semigroup in a variation of parameters formula to solve (2.5). However, the analysis of §§ 2 and 3 can with care be extended to cylindrical domains; most of the technicalities are contained in [10], and since we need $p \geqq 3$ and a uniqueness result analogous to Theorem 3.3 to obtain (2.36), there seems to be little advantage to us in the semigroup method. (The semigroup solution would satisfy (3.6), and hence coincide with $\tilde{u}$, so we would obtain the same solution by that approach. If we decompose L_0 into a sum of a first-order operator and the second-order operator generated by $\nabla\xi \cdot \mu(x, \xi)\nabla\xi$, we could recover the positivity of the semigroup by noting that each operator separately generates a positivity preserving semigroup and apply the Trotter product formula.)

In summary, we see that there are various possible formulations of L^p weak solutions to (2.5) and various ways of producing them. Our choice of approach is based on our interest in keeping as much information as possible about boundary behavior of our solution. Under sufficiently strong hypotheses, the various forms of weak solution all agree; however, without some added hypotheses uniqueness may fail.

4. General boundary conditions. In this section, we shall extend the existence and uniqueness theory developed in §§ 2 and 3 so as to include more general boundary conditions encountered in transport problems. The very nature of the *transport* problem implies that such a boundary condition relates distributions defined only on the characteristic parts of the boundary (in our case, D^+ and D^-):

$$u^- = Ku^+ + g \quad \text{on } D^-,$$

where K is a bounded operator,

$$K : L^p(D^+, d\nu^+) \to L^p(D^-, d\nu^-),$$

and, as before, g accounts for the autonomous boundary source. The distribution on the noncharacteristic part, Σ_3, is set to zero as in (2.5). Our main result is summarized in Theorem 4.1.

THEOREM 4.1. *Consider the following problem*:

$$\begin{aligned} Lu &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Sigma_3, \\ u^- &= Ku^+ + g && \text{on } D^- \end{aligned} \tag{4.1}$$

where Ω *satisfies the hypotheses of Lemma* 3.1, $f \in L^p(\Omega)$, $g \in L^p(D^-, d\nu^-)$, $3 \leqq p < \infty$, $\lambda \geqq 1$, *and* $\|K\| < 1$. *Then there exists a unique solution of* (4.1) *in* E_p *in the sense of* (2.7).

Proof. When $K = 0$ we may apply Theorems 2.4 and 3.2 and obtain a unique weak solution $(u, u^+, u^-) \in E_p$ with $u^- = g$ which satisfies (2.14), that is,

$$\|u\|^p_{L^p(\Omega)} + \|u^+\|_{L^p(D^+, d\nu^+)} \leqq \|f\|^p_{L^p(\Omega)} + \|g\|^p_{L^p(D^-, d\nu^-)}.$$

Let us call this solution $(u, u^+, u^-) := (T_\lambda(f, g), T_\lambda(f, g)^+, T_\lambda(f, g)^-)$. Observe that T_λ, T_λ^+, and T_λ^- are linear in f and g and $T_\lambda(f, g)^- \equiv g$. Using the linearity of T_λ and T_λ^+ and the estimate (2.14) we have

$$
\begin{aligned}
&\| T_\lambda(f, 0)\|_{L^p(\Omega)} \leqq \|f\|_{L^p(\Omega)},\\
&\| T_\lambda(f, 0)^+\|_{L^p(D^+, d\nu^+)} \leqq \|f\|_{L^p(\Omega)},\\
&\| T_\lambda(0, g)\|_{L^p(\Omega)} \leqq \|g\|_{L^p(D^-, d\nu^-)},\\
&\| T_\lambda(0, g)^+\|_{L^p(D^+, d\nu^+)} \leqq \|g\|_{L^p(D^-, d\nu^-)}.
\end{aligned} \tag{4.2}
$$

Following [4], we seek a solution of (4.1) in the form

$$(u, u^+, u^-) = (T_\lambda(f, g^*), T_\lambda(f, g^*)^+, T_\lambda(f, g^*)^-) \tag{4.3}$$

where we must have g^* satisfying the fixed point equation

$$g^* \equiv T_\lambda(f, g^*)^- = KT_\lambda(f, g^*)^+ + g. \tag{4.4}$$

(Recall that $g^* = T_\lambda(f, g)^-$ by definition.) We may rewrite (4.4) as

$$g^* = KT_\lambda(f, 0)^+ + KT_\lambda(0, g^*)^+ + g$$

or

$$(1 - M_\lambda)g^* = g + KT_\lambda(f, 0)^+ \tag{4.5}$$

where

$$M_\lambda g^* := KT_\lambda(0, g^*)^+.$$

Observe that $M_\lambda : L^p(D^-, d\nu^-) \to L^p(D^-, d\nu^-)$. Equation (4.5) has a unique solution which may be expressed by expanding $(I - M_\lambda)^{-1}$ in a Neumann series $\sum_{k=0}^{\infty} M_\lambda^k$ provided $\|M_\lambda\| < 1$. But (4.2) implies

$$
\begin{aligned}
\|M_\lambda g^*\|_{L^p(D^-, d\nu^-)} &\leqq \|K\| \, \|T_\lambda(0, g^*)^+\|_{L^p(D^+, d\nu^+)}\\
&\leqq \|K\| \, \|g^*\|_{L^p(D^-, d\nu^-)}
\end{aligned}
$$

so that $\|M_\lambda\| \leqq \|K\| < 1$. Then (4.5) has a unique solution g^*, and hence (4.1) has the unique solution

$$(u, u^+, u^-) = (T_\lambda(f, g^*), T_\lambda(f, g^*)^+, g^*).$$

This argument is equivalent to using the contraction mapping theorem or the Picard iteration method. □

Remark. In fact, we could use the more refined estimate (2.32) to assert

$$\| T_\lambda(f, 0)\|_{L^p(\Omega)} \leqq \lambda^{-1/q} \|f\|_{L^p(\Omega)}$$

and

$$\| T_\lambda(f, 0)^+\|_{L^p(D^+, d\nu^+)} \leqq \lambda^{-1/q} \|f\|_{L^p(\Omega)}$$

where $1/p + 1/q = 1$; however, that estimate is not required for our analysis.

5. Hilbert space solutions. Seeing the difficulties with uniqueness in weak solutions in E^p, $1 < p < 3$, we turn our attention to a stronger notion of solution in a "Sobolev type" Hilbert space. This approach will give uniqueness in its class and enables us to compare our results with recent results of Degond and Mas-Gallic [7].

We construct a Hilbert space that reflects the degeneracy in our equation. This construction is similar to the construction in Oleinik and Radkevic [10, § 1.4]. Define a class of test functions

$$\mathcal{W}=\{v\in C^1(\bar{\Omega}_0)\colon v=0 \text{ on } \Sigma_3^0\}.$$

For $(u, v)\in\mathcal{W}$, define an inner product

$$(u, v)_{\mathcal{H}}=\int_{\Omega_0}\left(\sum_{i=1}^{N} u_{\xi_i}v_{\xi_i}+uv\right)dx\,d\xi+\int_{D_0^+} u^+v^+\,d\nu_0^+ +\int_{D_0^-} u^-v^-\,d\nu_0^-$$

and define the Hilbert space $\mathcal{H}$ to be the closure of $\mathcal{W}$ with respect to the norm from the above inner product,

$$\|u\|_{\mathcal{H}}^2=(u, u)_{\mathcal{H}} \quad \text{for } u\in\mathcal{W}.$$

Define the bilinear form, for $u, v\in\mathcal{W}$,

$$B(u, v)=\int_{\Omega_0}\left[-\sum_{i=1}^{N} v_{\xi_i}u_{\xi_i}+u\left(-\sum_{i=1}^{N}\xi_i v_{\xi_i}+\sum_{i=1}^{N}\xi_i v_{x_i}\right)-(\lambda+1)uv\right]dx\,d\xi$$
$$-\int_{D_0^+} u^+v^+\,d\nu^+.$$

The definition of $B(u, v)$ may be extended to all functions $u\in\mathcal{H}$, $v\in\mathcal{W}$, and

$$|B(u, v)|\leqq C\left[\int_{\Omega_0}\left[\sum_{i=1}^{N}(v_{\xi_i}^2+v_{x_i}^2)+v^2\right]dx\,d\xi+\int_{D_0^+}(v^+)^2\,d\nu_0^+\right]^{1/2}\|u\|_{\mathcal{H}}$$

where C depends on the coefficients of the operator L_0. For fixed $v\in\mathcal{W}$, $B(u, v)$ is a bounded linear functional on $\mathcal{H}$.

DEFINITION. For $f\in L^2(\Omega_0)$, $g\in L^2(D_0^-, d\nu_0^-)$, a function u in $\mathcal{H}$ is a *weak solution in* $\mathcal{H}$ of

$$\begin{aligned} &L_0u=f \quad \text{in } \Omega_0,\\ &u^-=g \quad \text{on } D_0^-,\\ &u=0 \quad \text{on } \Sigma_3^0 \end{aligned} \tag{5.1}$$

if for all $v\in\mathcal{W}$,

$$B(u, v)=\int_{\Omega_0} fv\,dx\,d\xi-\int_{D_0^-} v^-g\,d\nu_0^-.$$

THEOREM 5.1. *Suppose* $\lambda>-\frac{1}{2}$, $f\in L^2(\Omega_0)$, $g\in L^2(D_0^-, d\nu_0^-)$; *then there exists a unique weak solution in* $\mathcal{H}$ *of* (5.1).

Proof. By the Riesz Theorem on the representation of linear functionals in Hilbert space, there exists a linear operator T on $\mathcal{W}$ with range in $\mathcal{H}$ such that

$$B(u, v)=(u, T(v))_{\mathcal{H}}.$$

Since $\lambda>-\frac{1}{2}$, $B(v, v)$ is strongly coercive,

$$|B(v, v)|=\int_{\Omega_0}\left[\sum_{i=1}^{N} v_{\xi_i}^2+\left(\lambda+\frac{1}{2}\right)v^2\right]dx\,d\xi+\frac{1}{2}\int_{D_0^+}(v^+)^2\,d\nu_0^+ +\frac{1}{2}\int_{D_0^-}(v^-)^2\,d\nu_0^-$$
$$\geqq\alpha\|v\|_{\mathcal{H}}^2, \qquad \alpha>0 \quad \text{for } v\in\mathcal{W}.$$

This coercivity implies

$$\|v\|_{\mathcal{H}}\leqq 1/\alpha\,\|T(v)\|_{\mathcal{H}}$$

and T is a one-to-one map. Denote by $\mathcal{H}_1$ the closure of the range of T in the norm of $\mathcal{H}$. Then

$$\int_{\Omega_0} vf\,dx\,d\xi - \int_{D_0^-} v^- g\,dv_0^-$$

is a continuous linear functional on $\mathcal{H}_1$. By the Riesz Theorem, there exists u in $\mathcal{H}_1$ such that

$$\int_{\Omega_0} vf\,dx\,d\xi - \int_{D_0^-} v^- g\,dv_0^- = (u, T(v))_{\mathcal{H}} = B(u, v) \quad \text{for all } v \in \mathcal{W}.$$

Thus u is a weak solution in $\mathcal{H}$.

By approximating u by a sequence $\{u_n\}$ in $\mathcal{W}$ with

$$\|u_n - u\|_{\mathcal{H}} \to 0 \quad \text{and} \quad B(u_n, v) \to B(u, v) \quad \text{for all } v \in \mathcal{W},$$

and integrating by parts on $B(u_n, v)$, we obtain

$$\int_{\Omega_0} (L_0^* v) u\,dx\,d\xi - \int_{D_0^+} v^+ u^+\,dv_0^+ = \int_{\Omega_0} vf\,dx\,d\xi + \int_{D_0^-} v^- g\,dv_0^- \tag{5.2}$$

$$\text{for all } v \in C^2(\bar{\Omega}), \quad v = 0 \text{ on } \Sigma_3^0.$$

We can use (5.2) to show the uniqueness of our weak solution in $\mathcal{H}$. Suppose u, $\tilde{u}$ are two weak solutions in $\mathcal{H}$, then

$$\int_{\Omega_0} (L_0^* v)(u - \tilde{u})\,dx\,d\xi = 0 \quad \text{for all } v \in C^2(\bar{\Omega}_0), \tag{5.3}$$

$$v = 0 \quad \text{on } \Sigma_3^0 \cup D_0^+.$$

By a result of Phillips and Sarason [10, Thm. 1.6.7], condition (5.3) implies $u = \tilde{u}$ almost everywhere. □

Remark. If $f \in L^p(\Omega_0)$, $g \in L^p(D_0^-, dv_0^-)$, $p \geqq 2$, and if $u \in \mathcal{H} \cap E_0^p$, condition (5.2) would imply that u would be a weak solution in E_0^p. But u is not necessarily in E_0^p. We can conclude that if $g = 0$ on D_0^-, $u \in L^p(\Omega_0)$, and then u is a weak solution in L^p sense, without the E_0^p requirement of L^p traces.

We now discuss another Hilbert space solution analyzed in a recent paper of Degond and Mas-Gallic [7]. They consider a one-dimensional electron scattering equation of the following form:

$$\sigma \frac{\partial}{\partial \xi}\left((1-\xi^2)\frac{\partial u}{\partial \xi}\right) - \xi \frac{\partial u}{\partial x} - \lambda u = f(x, \xi) \quad \text{for } (x, \xi) \in [0, L] \times [-1, 1],$$

$$u(0, \xi) = u_0(\xi) \quad \text{for } \xi \geqq 0, \tag{5.4}$$

$$u(L, \xi) = u_L(\xi) \quad \text{for } \xi \geqq 0, \quad \lambda > 0.$$

This particular solution Hilbert space Y is defined by

$$V = \left\{ \varphi \in L^2(-1, 1) : \sqrt{1-\xi^2}\,\frac{\partial \varphi}{\partial \xi} \in L^2(-1, 1) \right\},$$

$$\|\varphi\|_V = \int_{-1}^{1} \varphi^2(\xi)\,d\xi + \int_{-1}^{1} (1-\xi^2)\left(\frac{d\varphi}{d\xi}\right)^2 d\xi,$$

$$X = L^2([0, L], V) \qquad Y = \left\{ u \in X : \xi \frac{\partial u}{\partial x} \in X^* \right\}$$

where X^* is the dual space of X. Degond and Mas-Gallic obtain an existence and uniqueness result in Y. Their assumptions,

$$u \in L^2 \quad \text{and} \quad \xi \frac{\partial u}{\partial x} \in L^2,$$

provide L^2 traces on $D_0^+ \cup D_0^-$ by first-order methods. (See, for instance, [4].) Notice that the sides, $\xi = \pm 1$, are Σ_0 in our notation, and traces are not needed there. The combination of operator and domain is such that a priori estimates derived from (5.5) are all that are necessary; their a priori estimates can be obtained by only taking into account the first-order terms.

The weighted first derivative regularity in Degond and Mas-Gallic's paper is stronger than the regularity required in our case. This additional regularity yields also a strongly continuous semigroup of contractions on L^2 associated with the operator in (5.4).

Notice their second-order coefficient

$$\mu(\xi) = 1 - \xi^2,$$

becomes zero on the side boundaries, $\xi = \pm 1$. In § 2, we considered only strictly positive functions μ. The theory of equations with nonnegative characteristic forms in Oleinik and Radkevic [10] covers the case of such nonnegative coefficients μ. But including such coefficients involves redefining the characteristic parts of the boundary, which did not suit our goal here.

Acknowledgments. At an earlier stage of the preparation of the paper, Suzanne Lenhart and Vladimir Protopopescu benefitted from helpful comments from Professor R. Beals.

REFERENCES

[1] R. BEALS, *An abstract treatment of some forward-backward problems of transport and scattering*, J. Math. Anal. Appl., 34 (1979), pp. 1-20.

[2] ———, *Indefinite Sturm-Liouville problems and half range completeness*, J. Differential Equations, 56 (1985), pp. 391-407.

[3] R. BEALS AND V. PROTOPOPESCU, *Half range completeness for the Fokker-Planck equations*, J. Statist. Phys., 32 (1983), pp. 565-584.

[4] ———, *Abstract time-dependent transport equations*, J. Math. Anal. Appl., 12 (1987), pp. 370-405.

[5] H. A. BETHE, M. E. ROSE, AND L. P. SMITH, *The multiple scattering of electrons*, Proc. Amer. Philos. Soc., 78 (1938), pp. 573-585.

[6] S. CHANDRASEKHAR, *Stochastic problems in physics and astronomy*, Rev. Modern Phys., 15 (1943), pp. 1-89.

[7] P. DEGOND AND S. MAS-GALLIC, *Existence of solutions and diffusion approximation for a model Fokker-Planck equation*, Transport Theory Statist. Phys., to appear.

[8] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[9] T. G. GENCEV, *Ultraparabolic equations*, Soviet Math. Dokl., 4 (1963), pp. 979-982.

[10] O. A. OLEINIK AND E. V. RADKEVIC, *Second Order Equations with Nonnegative Characteristic Form*, American Mathematical Society, Providence, RI; Plenum Press, New York, 1973.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

004

ON EXISTENCE OF SOLUTIONS FOR SEVERAL CLASSES OF FREE BOUNDARY PROBLEMS*

PHILIP KORMAN†

Dedicated to Ken Meyer on the occasion of his 50th birthday.

Abstract. Some general techniques are developed for treating nonstandard nonlinear elliptic problems arising when a so-called domain perturbation method is applied to free boundary problems. Our results are applied to two model problems from fluid mechanics.

Key words. free boundary problems, domain perturbation method, existence and uniqueness of solutions

AMS(MOS) subject classification. 35J60

1. Introduction. Our goal is to explore a general approach to free boundary problems, based on the so-called domain perturbation method. Using this method we get a solution, which is a perturbation of a known one, by mapping (nonconformally) the unknown fluid domain back to the known domain for an unperturbed solution. The linear equations of motion then get transformed to a fully nonlinear system, but since nonlinearities are small, it can usually be treated by the contractive mapping argument. The earliest reference that we know for the domain perturbation method is Joseph [6]. It was then used by Shinbrot [10] to prove the existence of double-periodic water waves in three dimensions. Our work was motivated by that paper.

Rather than present our results in general form, we prefer to consider two model problems, whose treatment illustrates how one should approach various possibilities, and which are of considerable independent interest.

The nonlinear elliptic problems obtained by the domain perturbation method can be either coercive or noncoercive (here "coercive" means that the problem satisfies the Lopatinski–Schapiro condition at all points of the boundary). For the coercive problems we outline an approach using the Schauder-type estimates of Agmon, Douglis, and Nirenberg [1], and present it for the model Problem I. For noncoercive problems there are no Schauder's estimates. Estimates in the Sobolev spaces (which are available for Problems I and II) cannot be used, because of loss of smoothness when taking traces. For the model Problem II we present the second approach based on Λ^m spaces, which are defined and studied below. Similar spaces were used by Shinbrot [10]; however, ours have several advantages: it is easier to establish their properties, the proofs of the estimates for $m>2$ are more transparent, and finally they seem to be more natural. Problem II leads to a coercive problem, so that the first approach based on Schauder's estimates can be used as well. In § 6 we present an example of a physically significant problem, leading to a noncoercive problem which can be solved only by the second approach. We proceed to describe our model problems.

Problem I. Let $x \in R^n$. Given a 2π periodic in each variable x_i function $B(x)$ (bottom), find the functions $u(x, y)$, $H(x)$, 2π periodic in each x_i, such that

$$\begin{aligned} &u=0, \quad \frac{\partial u}{\partial n}=-1, \quad y=H(x), \\ &\Delta u=0, \qquad B(x)<y<H(x), \\ &u=1, \qquad y=B(x), \end{aligned} \tag{1.1}$$

* Received by the editors December 22, 1986; accepted for publication (in revised form) June 8, 1987.
† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221.

where $\partial/\partial n$ is the outward normal derivative.

Problem II. Let $r = B(\theta)$ be a closed curve in the plane. Find another closed curve $r = H(\theta)$ outside of $B(\theta)$, and a function u on the closed region between $B(\theta)$ and $H(\theta)$ with

$$\begin{aligned} &u = 0, \quad \frac{\partial u}{\partial n} = -1, \quad r = H(\theta), \\ &\Delta u = 0, \qquad B(\theta) < r < H(\theta), \\ &u = 1, \qquad r = B(\theta). \end{aligned} \tag{1.2}$$

For Problem I we start with flat bottom $B = 0$, and the corresponding solution $H = 1$ and $u = 1 - y$. Then for small bottoms $B = \varepsilon b(x)$ we are looking for the solution in the form

$$H = 1 + \varepsilon h(x), \qquad u = 1 - y + \varepsilon v(x, y), \tag{1.3}$$

and show existence if ε is sufficiently small. We use the change of variables $(x, y) \to (x, y')$, $y' = (y - \varepsilon b)/(1 + \varepsilon h - \varepsilon b)$, to transform the unknown domain onto a fixed one, $0 \leqq y' \leqq 1$.

For Problem II notice that if $B(\theta) = 1$ then the solution is $H(r) = h_0$, $u = -h_0 \log r + 1$, where $h_0 \simeq 1.76$ is defined by $h_0 \log h_0 = 1$, and (r, θ) are the polar coordinates. Then we assume that $B = 1 + \varepsilon b(\theta)$, and look for the solution in the form

$$H = h_0 + \varepsilon h(\theta), \qquad u = -h_0 \log r + 1 + \varepsilon v(r, \theta). \tag{1.4}$$

We show the existence of such a solution for ε sufficiently small.

Problem II was considered by Hamilton [5] (and also earlier by Schaeffer [9] and Acker [11]). Hamilton proved that for every smooth convex curve B there exists a unique solution to Problem II (the curve $H(\theta)$ is also smooth and convex, and u is smooth). His result is strictly two-dimensional, since conformal mappings were used to derive a priori estimates. Our existence result complements Hamilton's in that we do not require the curve $r = B(\theta)$ to be convex. In three dimensions we were unable so far to carry out a similar approach, because of the singularities in the Laplace operator in spherical coordinates.

We wish to stress the generality of our approach. It can be used to attack problems with boundary conditions of arbitrary order and variable coefficients, and with nonlinear equations of motion. In contrast, a more common variational approach (see, e.g., [2, Chap. 3]) is rather restricted (but it is a global method).

Finally, we mention that Problems I and II have an interesting physical interpretation. For Problem II it is described in [5, p. 215], so that we present a similar interpretation for Problem I (with similar deficiency as mentioned in [5]). We consider fluid occupying the half space $y > 0$, $x \in R^n$, and assume there is a stream flowing over the periodic bottom $y = B(x)$. The fluid is assumed to be perfect with unit density, and at rest outside a free surface boundary, so that there is a velocity jump at the free surface. Let u be the stream potential. By choosing units of length and time we can make the velocity on free surface $y = H(x)$ and circulation equal to one. This leads us to Problem I.

2. Preliminary results. Let $x = (x_1, \cdots, x_n)$, $j = (j_1, \cdots, j_n)$, $n \geqq 1$. Let the function $u = u(x, y)$ be 2π periodic in each variable x_i, $i = 1, \cdots, n$, $0 \leqq y \leqq 1$; $u(x, y) = \sum_{j=-\infty}^{\infty} u_j(y)\, e^{ij \cdot x}$. Define the norms $\|u(x, y)\|_0 = \sum_{j=-\infty}^{\infty} \max_{0 \leqq y \leqq 1} |u_j(y)|$, $\|u\|_m = \sum_{|\alpha| \leqq m} \|D^\alpha u\|_0$, where D^α is a mixed partial in x and y, m = integer $\geqq 1$. Denote

$D=[0,2\pi]^n$, $V=D\times[0,1]$. Let $\Lambda^m(V)$ be the closure of trigonometric polynomials of the form $\sum u_j(y)\,e^{ij\cdot x}$, $u_j(y)\in C^\infty[0,1]$, with respect to the norm $\|\cdot\|_m$. Clearly $\Lambda^m(V)$ are Banach spaces with $\|u\|_m\leqq\|u\|_n$ if $m\leqq n$. The space $\Lambda^m(D)$ is defined in the same way for functions independent of y. The norm on $\Lambda^m(D)$ is denoted by $|||\cdot|||_m$. If we are given a function of polar coordinates in the plane, $u=u(r,\theta)=\sum_{n=-\infty}^{\infty}u_n(r)\,e^{in\theta}$ on an annulus $1\leqq r\leqq h_0$, $h_0=$ constant, then as before $\|u\|_0=\sum_{n=-\infty}^{\infty}\max_{1\leqq r\leqq h_0}|u_n(r)|$, and $\|u\|_m=\sum_{|\alpha|\leqq m}\|D^\alpha u\|_0$ where D^α is a mixed partial in r and θ. This time domain V is defined by $1\leqq r\leqq h_0$, $0\leqq\theta\leqq 2\pi$ domain D by $r=h_0$, $0\leqq\theta\leqq 2\pi$.

We write c for all positive constants independent of unknown functions. We write $f=f(D^2v)$ when f depends on the function v and all its partial derivatives or orders one and two.

LEMMA 2.1. *Let* $u,v\in\Lambda^m$. *Then* $uv\in\Lambda^m$, *and* $\|uv\|_m\leqq c_m\|u\|_m\|v\|_m$, $c_m=$ const $(c_0=1)$.

Proof. Let $u=\sum_{j=-\infty}^{\infty}u_j\,e^{ij\cdot x}$, $v=\sum_{k=-\infty}^{\infty}v_k\,e^{ik\cdot x}$. Then

$$\|uv\|_0=\sum_\gamma\max_y\left|\sum_\delta u_{\gamma-\delta}v_\delta\right|\leqq\sum_\delta\max_y|v_\delta|\sum_\gamma\max_y|u_{\gamma-\delta}|$$
$$=\|u\|_0\|v\|_0.$$

For $m\geqq 1$ we get

$$\|uv\|_m=\sum_{|\alpha|\leqq m}\|D^\alpha uv\|=\sum_{|\alpha|\leqq m}\sum_{0\leqq\beta\leqq\alpha}c_\beta\|D^\beta u\|_0\|D^{\alpha-\beta}v\|_0$$
$$\leqq c_m\|u\|_m\|v\|_m.$$

COROLLARIES. (i) $\|g_1\cdots g_p\|_m\leqq c_m^{p-1}\|g_1\|_m\cdots\|g_p\|_m$.

(ii) $\|g^p\|_m\leqq c_m^{p-1}\|g\|_m^p$.

LEMMA 2.2. *Let* B *be a ball in* R^p *centered at the origin,* $f(x_1,\cdots,x_p)$: $B\to R^1$ *be a real analytic function. Let* g *be a vector function on* V, $g=(g_1,\cdots,g_p)$, $\|g\|_m\equiv\sum_{i=1}^p\|g_i\|_m=r$. *Assume that* r *is sufficiently small. Then* $f(g_1,\cdots,g_p)\in\Lambda^m$, *and*

$$\|f(g)\|_m\leqq c_0+c_1(r),$$

where $c_0=\text{const}>0$, c_1 *is analytic function of* r, *depending only on* f *and* r, *and* $c_1(0)=0$.

Proof. Let $f=\sum_{|\alpha|\geqq 0}f_\alpha x^\alpha$ for $x=(x_1,\cdots,x_p)\in B$, $f(g)=\sum_{|\alpha|\geqq 0}f_\alpha g_\alpha$. Then by Lemma 2.1

$$\|f(g)\|_m\leqq\sum_{|\alpha|\geqq 0}|f_\alpha|\,\|g^\alpha\|_m\leqq\sum_{|\alpha|\geqq 0}|f_\alpha|c_m^{|\alpha|-1}r^{|\alpha|},$$

which is easily seen to be a convergent series for r sufficiently small.

LEMMA 2.3. $\Lambda^m(V)(\Lambda^m(D))$ *is boundedly imbedded in* $C^m(V)(C^m(D))$.

Proof. Since for any multi-index α, $|\alpha|=m$, $\max_{x,y}|D^\alpha u|\leqq\|u\|_m$ the proof follows.

By $|\cdot|_{m+\alpha}$ we denote the norm in the space $C^{m+\alpha}(V)$, $m=\text{integer}\geqq 0$, $0<\alpha<1$ (see, e.g., [1] for the definition).

3. Transformation to a fixed domain. For Problem I we suppose that $B(x)=\varepsilon b(x)$, and look for solution in the form (1.3). Notice that on $y=H(x)=1+\varepsilon h(x)$

$$\frac{\partial u}{\partial n}=\nabla u\cdot n=\frac{-\varepsilon\sum_{i=1}^n u_ih_i+u_y}{\sqrt{1+\varepsilon^2|\nabla h|^2}}.$$

Substituting this and (1.3) into (1.1) we get

$$v(x,y)=h(x),\qquad\frac{-\varepsilon^2\sum_{i=1}^n v_ih_i-1+\varepsilon v_y}{\sqrt{1+\varepsilon^2|\nabla h|^2}}=-1,\qquad y=1+\varepsilon h,$$

$$\Delta v=0,\qquad\varepsilon b(x)<y<1+\varepsilon h(x),\tag{3.1}$$
$$v=b(x),\qquad y=\varepsilon b(x).$$

The change of variables $(x, y) \to (x', y')$ defined by

$$x_i' = x_i, \qquad i = 1, \cdots, n,$$

$$y' = \frac{y - \varepsilon b}{1 + \varepsilon(h - b)} \equiv y d(x) + e(x),$$

$$\left(d(x) = \frac{1}{1 + \varepsilon(h - b)}, \; e(x) = -\frac{\varepsilon b}{1 + \varepsilon(h - b)} \right)$$

will transform the unknown fluid domain onto $0 \leqq y' \leqq 1$.

By a straightforward calculation, the problem (3.1) will transform as follows (we drop primes for the independent variables)

$$\begin{aligned} & v(x, 1) = h(x), \\ & v_y = \varepsilon g(\varepsilon, Dv, Dh, Db), && y = 1, \\ & \Delta v = \varepsilon f(\varepsilon, D^2 v, D^2 h, D^2 b), && 0 < y < 1, \\ & v = b(x), && y = 0. \end{aligned} \tag{3.2}$$

Here

$$\varepsilon g = \frac{1 - \sqrt{1 + \varepsilon^2 |Dh|^2} + \varepsilon^2 \sum_{i=1}^{n} [v_{x_i} + v_y (Y d_{x_i} + e_{x_i})] h_{x_i}}{\varepsilon d(x)}, \tag{3.3}$$

where $Y = y(1 + \varepsilon h - \varepsilon b) + \varepsilon b$, and

$$\begin{aligned} -\varepsilon f = (d^2 - 1) v_{yy} + \sum_{i=1}^{n} & [2 v_{x_i y} (Y d_{x_i} + e_{x_i}) + v_{yy} (Y d_{x_i} + e_{x_i})^2 \\ & + v_y (Y d_{x_i x_i} + e_{x_i x_i})]. \end{aligned} \tag{3.4}$$

We easily see that the functions f and g are analytic in their arguments for small ε.

For Problem II we suppose that $B = 1 + \varepsilon b(\theta)$, and look for a solution in the form (1.4). By an elementary computation on $r = H(\theta) = h_0 + \varepsilon h(\theta)$ we have

$$\frac{\partial u}{\partial n} = \frac{r}{\sqrt{r^2 + \varepsilon^2 h'^2}} u_r - \frac{\varepsilon h'}{r \sqrt{r^2 + \varepsilon^2 h'^2}} u_\theta. \tag{3.5}$$

Using this formula and (1.4) in (1.2) we get

$$\begin{aligned} & v = \frac{-1 + h_0 \log r}{\varepsilon}, && r = h_0 + \varepsilon h, \\ & \frac{r}{\sqrt{r^2 + \varepsilon^2 h'^2}} \left(-\frac{h_0}{r} + \varepsilon v_r \right) - \frac{\varepsilon^2 h'}{r \sqrt{r^2 + \varepsilon^2 h'^2}} v_\theta = -1, && r = h_0 + \varepsilon h, \\ & \Delta v = 0, && 1 + \varepsilon b(\theta) < r < h_0 + \varepsilon h(\theta), \\ & v = \frac{h_0 \log r}{\varepsilon}, && r = 1 + \varepsilon b(\theta). \end{aligned} \tag{3.6}$$

The change of variables $(r, \theta) \to (r', \theta')$

$$r' = (h_0 - 1) \frac{r - 1 - \varepsilon b}{\varepsilon(h - b) + h_0 - 1} + 1, \qquad \theta' = \theta$$

maps the fluid domain $1+\varepsilon b \leqq r \leqq h_0+\varepsilon h$ onto the annulus $1 \leqq r' \leqq h_0$. The problem (3.5) will transform as follows (dropping the primes)

$$
\begin{aligned}
&v = h + \varepsilon r(\varepsilon, h), && r = h_0,\\
&v_r = -\frac{h}{h_0} + \varepsilon g(\varepsilon, Dv, Dh, Db), && r = h_0,\\
&\Delta v = \varepsilon f(\varepsilon, D^2 v, D^2 h, D^2 b), && 1 < r < h_0,\\
&v = h_0 b + \varepsilon q(\varepsilon, b), && r = 1.
\end{aligned} \tag{3.7}
$$

Here

$$
\varepsilon r = \frac{h_0 \log (h_0+\varepsilon h) - 1}{\varepsilon} - h,
$$

$$
\varepsilon g = \frac{\varepsilon}{(h_0+\varepsilon h)^2 p}\left[v_\theta h' + v_r(Rp_\theta + q_\theta)h' - h^2 - (h_0+\varepsilon h)^2 r_1\right] + \frac{1}{\varepsilon}\left(\frac{h}{h_0} - \frac{h_0 h}{r^2 p}\right)
$$

with

$$
p = \frac{h_0 - 1}{\varepsilon(h-b)+h_0-1}, \qquad q = -\frac{(h_0-1)(1+\varepsilon b)}{\varepsilon(h-b)+h_0+1} + 1,
$$

$$
R = \frac{r-q}{p}, \qquad r_1 = \frac{1}{\varepsilon^2}\left(\sqrt{1+\varepsilon^2\frac{h'^2}{R^2}} - 1\right),
$$

$$
-\varepsilon f = (p^2-1)v_{rr} + \left(\frac{p}{R} - \frac{1}{r}\right)v_r + \left(\frac{1}{R^2} - \frac{1}{r^2}\right)v_{\theta\theta}
$$

$$
+ \frac{1}{R^2}\left[2v_{\theta r}(Rp_\theta + q_\theta) + v_{rr}(Rp_\theta + q_\theta)^2 + v_r(Rp_{\theta\theta} + q_{\theta\theta})\right],
$$

$$
\varepsilon q = \frac{h_0}{\varepsilon}(\log(1+\varepsilon b) - \varepsilon b).
$$

We verify that the functions r, g, f, q are analytic in their arguments for small ε.

4. A priori estimates for the linear problem. Consider the problem ($x \in R^n$)

$$
\begin{aligned}
&u_y = g(x), && y = 1,\\
&\Delta u = f(x, y), && 0 < y < 1,\\
&u = b(x), && y = 0,
\end{aligned} \tag{4.1}
$$

where f, g, and b are given functions, 2π periodic in each variable x_i, $i = 1, \cdots, n$.

LEMMA 4.1. *Assume that $f \in C^\alpha(V)$, $g \in C^{1+\alpha}(D)$. Then* (4.1) *has a unique 2π periodic in each x_i solution, and*

$$
|u|_{2+\alpha} \leqq c(|f|_\alpha + |g|_{1+\alpha} + |b|_{2+\alpha}). \tag{4.2}
$$

Proof. Existence of solutions follows by elementary Fourier analysis, uniqueness from the estimate

$$
|u|_0 \leqq c(|f|_0 + |g|_0 + |b|_0), \tag{4.3}
$$

which easily follows by the maximum principle. It remains to show how one adapts Schauder's estimates for our problem (4.1). Redefine f, g, b as functions of compact

support outside $0 \leqq x_i \leqq 2\pi$, $i = 1, \cdots, n$, and call the extensions $\bar{f}$, $\bar{g}$, $\bar{b}$, respectively. Clearly, this can be done with say $|\bar{f}|_\alpha \leqq 2|f|_\alpha$, $|\bar{g}|_\alpha \leqq 2|g|_\alpha$, $|\bar{b}|_\alpha \leqq 2|b|_\alpha$. Let $\xi_1(y)$, $\xi_2(y)$ be C^∞ functions on $[0, 1]$, such that $\xi_1 \equiv 1$ near $y = 1$ and $\xi_1 \equiv 0$ near $y = 0$, and $\xi_2 = 1 - \xi_1$. Write $u = \xi_1 u + \xi_2 u \equiv u_1 + u_2$. Multiplying (4.1) by ξ_1 and ξ_2, we easily get

$$
\begin{aligned}
& u_{1y} = \bar{g}, && y = 1, \\
& \Delta u_1 = \xi_1'' u + 2\xi_1' u_y + \xi_1 \bar{f}, && -\infty < y < 1,
\end{aligned} \tag{4.4}
$$

$$
\begin{aligned}
& \Delta u_2 = \xi_2'' u + 2\xi_2' u_y + \xi_2 \bar{f}, && 0 < y < \infty, \\
& u_2 = \bar{b}(x), && y = 0.
\end{aligned} \tag{4.5}
$$

Using usual Schauder's estimates (see [1, Thm. 7.3]), we get (for arbitrary small ε)

$$
\begin{aligned}
|u|_{2+\alpha} &\leqq c\left(\sum_{i=1}^{2} |\xi_i'' u + 2\xi_i' u_y + \xi_i \bar{f}|_\alpha + |\bar{g}|_{1+\alpha} + |\bar{b}|_{2+\alpha} \right) \\
&\leqq c(\varepsilon |u|_{2+\alpha} + c_\varepsilon |u|_0 + |f|_\alpha + |g|_{1+\alpha} + |b|_{2+\alpha}),
\end{aligned}
$$

and by (4.3) the lemma follows.

Next, in the plane (r, θ) consider the problem ($h_0 \log h_0 = 1$)

$$
\begin{aligned}
& u_r + \sigma u = g(\theta), && r = h_0 \quad (\sigma = \text{const} \geqq 0), \\
& \Delta u = f(r, \theta), && 1 < r < h_0, \\
& u = b(\theta), && r = 1.
\end{aligned} \tag{4.6}
$$

Here, g, f, b are given functions 2π periodic in θ.

LEMMA 4.2. *Assume* $b \in \Lambda^{m+2}$, $f \in \Lambda^m$, $g \in \Lambda^{m+1}$, $m \geqq 0$. *Then* (4.6) *has a unique solution and*

$$
\|u\|_{m+2} + \overline{\|u\|}_{m+2} \leqq c(\|f\|_m + \|g\|_{m+1} + \|b\|_{m+2}). \tag{4.7}
$$

Proof. Express $f = \sum_{n=-\infty}^{\infty} f_n(r)\, e^{in\theta}$, $g = \sum_{n=-\infty}^{\infty} g_n\, e^{in\theta}$, $b = \sum_{n=-\infty}^{\infty} b_n\, e^{in\theta}$, $u = \sum_{n=-\infty}^{\infty} u_n(r)\, e^{in\theta}$. Substituting these into (4.6) and suppressing the subscript n (i.e., writing f for f_n, g for g_n, etc.) and letting $r = e^x$, we get

$$
u_{xx} - n^2 u = e^{2x} f(e^x), \quad u(0) = b, \quad \frac{1}{h_0} u_x(h_1) + \sigma u(h_1) = g, \tag{4.8}
$$

where $h_1 = \log h_0 \simeq 0.57$. Set $F(t) = e^{2t} f(e^t)$. The solution of (4.8) is

$$
u(x) = \gamma \sinh nx + b \cosh nx + \frac{1}{n} \int_0^x F(t) \sinh n(x - t)\, dt, \tag{4.9}
$$

where the constant γ is determined from

$$
\begin{aligned}
& \gamma A + b\left(\frac{n \sinh nh_1}{h_0} + \sigma \cosh nh_1 \right) \\
& \quad + \int_0^{h_1} F(t) \left[\frac{1}{h_0} \cosh n(h_1 - t) + \frac{\sigma}{n} \sinh n(h_1 - t) \right] dt = g.
\end{aligned} \tag{4.10}
$$

Here we denoted

$$
A = \frac{n \cosh nh_1}{h_0} + \sigma \sinh nh_1 \geqq cn\, e^{nh_1}. \tag{4.11}
$$

Multiplying (4.9) by A, using (4.10) and the standard identities for hyperbolic functions, we easily derive

$$
\begin{aligned}
Au(x) &= g \sinh nx + b\left(\frac{n}{h_0}\cosh n(h_1-x) + \sigma \sinh n(h_1-x)\right) \\
&\quad - \int_x^{h_1} F(t)\left[\frac{1}{h_0}\cosh n(h_1-t)\sinh nx + \frac{\sigma}{n}\sinh n(h_1-t)\sinh nx\right] dt \qquad (4.12)\\
&\quad - \int_0^{x} F(t)\left[\frac{1}{h_0}\sinh nt \cosh n(h_1-x) + \frac{\sigma}{n}\sinh nt \sinh n(h_1-x)\right] dt.
\end{aligned}
$$

Then in view of (4.11) we easily estimate

$$
\begin{aligned}
|u(r)| &\leqq c\left[\frac{|g|}{n} + |b| + \frac{1}{n}\int_x^{h_1} |F(t)|\, e^{n(x-t)}\, dt + \frac{1}{n}\int_0^{x} |F(t)|\, e^{n(t-x)}\, dt\right] \\
&\leqq c\left[\frac{|g|}{n} + |b| + \frac{1}{n}\sup_{1\leqq r\leqq h_0} |f(r)|\left(\frac{2}{n} - \frac{e^{n(x-h_1)}}{n} - \frac{e^{-nx}}{n}\right)\right] \qquad (4.13)\\
&\leqq c\left[\frac{|g|}{n} + |b| + \frac{1}{n^2}\max_{1\leqq r\leqq h_0} |f(r)|\right].
\end{aligned}
$$

Differentiating (4.9) and going through the same steps, we estimate

$$
|u'(r)| \leqq c\left[|g| + n|b| + \frac{1}{n}\max_{1\leqq r\leqq h_0} |f(r)|\right]. \qquad (4.14)
$$

Combining (4.13) with (4.14), and estimating $|u''(r)|$ from the equation, we conclude the estimate (4.7) with $m = 0$. The higher estimates are easily proved by induction.

An a priori estimate for (4.1) is given by the following lemma whose proof is similar to the above.

LEMMA 4.3. *Assume* $f \in \Lambda^m$, $g \in \Lambda^{m+1}$, $b \in \Lambda^{m+2}$, $m = \text{integer} \geqq 0$. *Then* (4.1) *has a unique* 2π *periodic in each* x_i *solution, and*

$$
\|u\|_{m+2} + \overline{\|u\|}_{m+2} \leqq c(\|f\|_m + \|g\|_{m+1} + \|b\|_{m+2}).
$$

5. Existence and uniqueness of solutions for Problems I and II.

THEOREM 5.1. *For* (3.2) *assume that* $b(x) \in C^{m+\alpha}(D)$, *and* $\varepsilon|b|_{m+\alpha}$ *is sufficiently small,* $m = \text{integer} \geqq 2$, $0 < \alpha < 1$. *Then there exists a pair of functions* $(v, h) \in C^{m+\alpha}(V) \times C^{m+\alpha}(D)$ *satisfying* (3.2).

Proof. Define a map T: $(w, k) \to (v, h)$ from $C^{m+\alpha}(V) \times C^{m+\alpha}(D)$ to itself by solving

$$
\begin{aligned}
v_y &= \varepsilon g(\varepsilon, Dw, Dk, Db), && y = 1,\\
\Delta v &= \varepsilon f(\varepsilon, D^2 w, D^2 k, D^2 b), && 0 < y < 1,\\
v &= b(x), && y = 0,
\end{aligned}
$$

and then computing $h(x) = v(x, 1)$. By Lemma 4.1 we easily conclude that the map T is well defined, takes a ball $|w|_{m+\alpha} + |k|_{m+\alpha} \leqq R$, with say $R = 2|b|_{m+\alpha}$, into itself, and is a contraction for $\varepsilon|b|_{m+\alpha}$ sufficiently small.

A similar proof could be given for Problem II. Instead, we give an existence proof in Λ^m spaces based on Lemma 4.2, which provides a more general approach, as will be seen in § 6.

THEOREM 5.2. *For the problem* (3.7) *assume that* $b(\theta) \in \Lambda^m$, *and* $\varepsilon \|b\|_m$ *is sufficiently small,* $m = \text{integer} \geqq 2$. *Then there exists a pair of functions* $(v, h) \in \Lambda^m(V) \times \Lambda^m(D)$ *satisfying* (3.6).

Proof. Substituting the first equation in (3.7) into the second we get

$$v_r + \frac{1}{h_0} v = \frac{\varepsilon}{h_0} r(\varepsilon, h) + \varepsilon g \equiv \varepsilon \bar{g}(\varepsilon, Dv, Dh, Db).$$

Next, we define a map $T: (w, k) \to (v, h)$ from $\Lambda^m(V) \times \Lambda^m(D)$ to itself by solving

$$\begin{aligned} &v_r + \frac{1}{h_0} v = \varepsilon \bar{g}(\varepsilon, Dw, Dk, Db), && r = h_0, \\ &\Delta v = \varepsilon f(\varepsilon, D^2 w, D^2 k, D^2 b), && 1 < r < h_0, \\ &v = h_0 b + \varepsilon q(\varepsilon, b), && r = 1, \end{aligned} \tag{5.1}$$

and then solving for $h(\theta)$ from

$$v(h_0, \theta) = h + \varepsilon r(\varepsilon, h) \left(h = \frac{h_0}{\varepsilon} (e^{\varepsilon v / h_0} - 1) \right). \tag{5.2}$$

Notice that the map T is well defined, i.e., it takes $\Lambda^m(V) \times \Lambda^m(D)$ into itself, provided $\varepsilon \|b\|_m$ is sufficiently small. Indeed, in view of the estimate (4.7) of Lemma 4.2, it suffices to show that $\bar{g} \in \Lambda^{m-1}(D)$, $f \in \Lambda^{m-2}(V)$. For this we use the special structure of f and $\bar{g}$. Indeed, consider $\bar{g}$. By Lemma 2.3, $\log(h_0 + \varepsilon h)$, p_θ, q_θ, $1/(h_0 + \varepsilon h)^2 \in \Lambda^{m-1}(D)$ for ε small (if smallness of $\varepsilon \|b\|_m$ comes from $\|b\|_m$, then work in small balls), and then by Lemma 2.2, $\bar{g} \in \Lambda^{m-1}(D)$. Similarly, $f \in \Lambda^{m-2}$ by Lemma 2.2.

Then one easily sees that map T takes the ball $\|w\|_m + \|k\|_m \leqq 2\|b\|_m$ into itself, and is a contraction.

Remark. A similar argument is valid for Problem I.

Next we prove uniqueness results, using techniques similar to [3] and [4].

THEOREM 5.3. *Problem* I *can have at most one solution* (*in the class of free surfaces satisfying interior sphere condition*).

Proof. Assume that this is not true, i.e., there are two solutions $(u(x, y), h(x))$ and $(\bar{u}(x, y), \bar{h}(x))$. By the maximum principle we conclude that $0 \leqq u \leqq 1$ for $b \leqq y \leqq h$, and $0 < u < 1$ for $b < y < h$, and also that h and $\bar{h}$ are different. Consider first the special case, when one free surface is above the other, touching at some point, say, $h(x) \geqq \bar{h}(x)$, $h(x_0) = \bar{h}(x_0)$. Consider $w = u - \bar{u}$ in the domain $b \leqq y \leqq \bar{h}$. Then $\Delta w = 0$, $w = 0$ for $y = b(x)$, $w \geqq 0$ for $y = \bar{h}(x)$ with $w(x_0) = 0$. Hence x_0 is a point of minimum for w. Since h and $\bar{h}$ have the same normal at x_0, by Hopf's lemma we have

$$0 > \frac{\partial w}{\partial n}(x_0) = \frac{\partial u}{\partial n}(x_0) - \frac{\partial \bar{u}}{\partial n}(x_0) = 0,$$

a contradiction.

Turning to the general case, we introduce translation of solution (u, h) downward, by considering

$$u_\tau(x, y) = u(x, y + \tau), \quad h_\tau(x) = h(x) - \tau, \quad \tau \geqq 0.$$

Clearly $\Delta u_\tau = 0$ for $b - \tau < y < h - \tau$. Choose $\tau = \tau_0$ so that $h_{\tau_0} \leqq \bar{h}$, and $h_{\tau_0}(x_0) = \bar{h}(x_0)$ for some x_0. (If h and $\bar{h}$ intersect, we translate either of two solutions, if $h > \bar{h}$ then translate (u, h).) Let $\bar{D} = \{x \mid x_0 \in D \text{ and } h_{\tau_0}(x) > b(x)\}$. By periodicity either $\bar{D} = (-\infty, \infty)$ or D is a bounded interval. In the first case consider $w = \bar{u} - u_{\tau_0}$ with $\bar{u}$, u_{τ_0}

restricted to $b(x) \leqq y \leqq h_{\tau_0}(x)$. Then $w > 0$ on $y = b(x)$, $w \geqq 0$ on $y = h_{\tau_0}(x)$, $w(x_0) = 0$, and as before we get contradiction at the point x_0. In the second case we consider the same w, with $\bar{u}$, u_{τ_0} restricted to $(x, y) = \{x \in \bar{D}, b(x) \leqq y \leqq h_{\tau_0}(x)\}$. Again we get the same contradiction at x_0.

THEOREM 5.4. *Problem* II *can have at most one solution* (*assuming* $h(\theta)$ *satisfying interior sphere condition*).

Proof. This time we introduce contraction of the solution by considering $u_a = u(ar, \theta) = u(ax, ay)$, $h_a = h/a$, $a > 1$. Clearly, $\Delta u_a = 0$. By contracting one of the free surfaces, until it is inside the other touching it at some point x_0, we get the same contradiction at x_0 as in the previous theorem. (Again, if the surfaces intersect, contract either one; if one is outside the other, contract the outside one. Also, notice that $\partial u_a / \partial n|_{h=h_a} = -a < -1$, as is clear from (3.5).)

6. General noncoercive problems. We discuss the problem ($x \in R^n$, $0 \leqq y \leqq 1$)

$$
\begin{aligned}
& u_y + \sum_{|\alpha| \leqq k} a_\alpha D^\alpha u = \varepsilon g(\varepsilon, D^k u, D^k h, D^k b), && y = 1, \\
& \Delta u = \varepsilon f(\varepsilon, D^2 u, D^2 h, D^2 b), && 0 < y < 1, \\
& u_y + \sum_{|\alpha| \leqq l} b_\alpha D^\alpha u = b(x), && y = 0.
\end{aligned} \tag{6.1}
$$

Here $h = u(x, 1)$, g, f, and b are given functions, 2π periodic in each variable x_i, $i = 1, \cdots, n$; $\alpha = (\alpha_1, \cdots, \alpha_n, 0)$, ε, a_α, b_α are constants, and k, l are integers whose magnitudes are not restricted. We are looking for 2π periodic in each x_i solution $u(x, y)$, assuming ε is small.

Solving a problem of type (6.1) was the key ingredient in solving Problems I and II, as well as in Shinbrot's proof of existence of water waves in three dimensions. If the boundary condition at $y = 1$ is coercive, i.e., satisfies the Lopatinski-Schapiro condition, then one should be able to prove existence based on Schauder's estimates, as we did for Problem I. In particular, in Shinbrot's paper one has the boundary operators (with $u = u(x, y, z)$) $u_y - \tau(u_{yxx} + u_{yzz}) + F u_{xx}$ at $y = 1$, and u_y at $y = 0$, which are both coercive. Hence, it appears that Schauder's estimates can be used, considerably simplifying the proof.

Using Λ^m spaces one can treat more general problems, including noncoercive ones. In particular, we have the following theorem, whose proof is similar to that of Theorem 5.2.

THEOREM 6.1. *Assume the following estimate for the problem* (6.1) (*with* $g = g(x)$, $f = f(x, y)$)

$$
\|u\|_m + \overline{\|u\|}_m \leqq c(\|f\|_{m-2} + \|g\|_{m-k} + \|b\|_m),
$$

with integer $m \geqq \max(2, k)$. *Assume that the functions* g *and* f *are analytic in their arguments and small if either* ε *or* $\|b\|_m$, $\|u\|_m$, $\|h\|_m$ *are sufficiently small. Then for* $\varepsilon \|b\|_m$ *sufficiently small the problem* (6.1) *has a solution.*

Example. In Shinbrot's water wave model assume that the surface tension $\tau = 0$, and the gravity is pointing up, i.e., g and $F = U^2/g$ are negative numbers, say $F = -f$, $f \geqq 0$. Assuming for simplicity $u = 0$ at $y = 0$, we consider the problem (different from the one in [10])

$$
\begin{aligned}
& u_y - f u_{xx} = \varepsilon g(\varepsilon, Du, Dh, Db), && y = 1, \\
& \Delta u = \varepsilon f(\varepsilon, D^2 u, D^2 h, D^2 b), && 0 < y < 1, \\
& u = 0, && y = 0.
\end{aligned} \tag{6.2}
$$

Here $u = u(x, y, z)$, $h = u(x, 1, z)$. The boundary condition at $y = 1$ is noncoercive (see [7]); hence Schauder's estimates are not valid for (6.2). However, by an argument similar to that of Lemma 4.2 we can estimate (with $g = g(x, z)$, $f = f(x, y, z)$)

$$\|u\|_{m+2} + \overline{\|u\|}_{m+2} \leq c(\|f\|_m + \|g\|_{m+1}),$$

and Theorem 6.1 applies, giving existence for (6.2). (We do not know any other way to prove existence for (6.2).)

Acknowledgments. I wish to thank A. Friedman, P. L. Lions, and S. Stojanovic for useful comments.

REFERENCES

[1] S. AGMON, A. DOUGLIS, AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, I, Comm. Pure Appl. Math., 12 (1959), pp. 623-727.

[2] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, John Wiley, New York, 1982.

[3] A. FRIEDMAN AND T. VOGEL, *Cavitational flow in a channel with oscillatory wall*, Nonlinear Anal. TMA, 7 (1983), pp. 1175-1192.

[4] D. GILBARG, *Uniqueness of axially symmetric flows with free boundaries*, Arch. Rational Mech. Anal., 1 (1952), pp. 309-320.

[5] R. HAMILTON, *The inverse function theorem of Nash and Moser*, Bull. Amer. Math. Soc., 7 (1982), pp. 65-222.

[6] D. JOSEPH, *Domain perturbations: the higher order theory of infinitesimal water waves*, Arch. Rational Mech. Anal., 51 (1973), pp. 295-303.

[7] P. KORMAN, *Existence of solutions for a class of nonlinear non-coercive problems*, Comm. Partial Differential Equations, 8 (1983), pp. 819-846.

[8] ———, *Existence of periodic solutions for a class of nonlinear problems*, Nonlinear Anal. TMA, 7 (1983), pp. 873-879.

[9] D. SCHAEFFER, *A stability theorem for the obstacle problem*, Adv. in Math., 17 (1975), pp. 34-47.

[10] M. SHINBROT, *Water waves over periodic bottoms in three dimensions*, J. Inst. Math. Applic., 25 (1980), pp. 367-385.

[11] A. ACKER, *A free boundary optimization problem* 1, SIAM J. Math. Anal., 9 (1978), pp. 1179-1191; II, 11 (1980), pp. 201-209.

SIAM J. MATH ANAL.
Vol. 19, No. 4, July 1988

005

GLOBAL BIFURCATION AND CONTINUATION IN THE PRESENCE OF SYMMETRY WITH AN APPLICATION TO SOLID MECHANICS*

TIMOTHY J. HEALEY†

Abstract. A group-theoretic approach to global bifurcation and continuation for one-parameter problems with symmetry is presented. The basic theme is the construction of a reduced problem, having solutions with specified symmetries, that can be analyzed by global or local techniques. A global analysis of a general class of reduced problems via well-established continuation techniques shows that symmetry is preserved on global continua of solutions. The approach is illustrated in the analysis of large post-buckling solutions of a nonlinearly elastic ring with $O(2)$ symmetry under uniform hydrostatic pressure, and yields several new results. Specific symmetries of global bifurcating solution branches are enumerated, which enables a detailed qualitative analysis.

Key words. symmetry, bifurcation, global, groups, solid mechanics, structures, post-buckling

AMS(MOS) subject classifications. 34, 58, 73

1. Introduction. Let B be a real Banach space, let Ω be an open, connected subset of $B\times\mathbb{R}$, and let $f:\Omega\to B$ be $(m\geqq 1)$-times continuously Fréchet differentiable. Consider a steady or static bifurcation problem of the form

$$f(x,\lambda)=0. \tag{1.1}$$

The goal is to determine the *solution set*

$$\Sigma\equiv\{(x,\lambda)\in\Omega: f(x,\lambda)=0\}. \tag{1.2}$$

Bifurcation problems often arise in the physical sciences for systems with symmetry. Suppose that (1.1) models such a system, characterized by a symmetry group $\mathcal{G}$. In such a case, (1.1) is usually *equivariant* under a specific representation (cf. Robert [1983]) T of $\mathcal{G}$ on B

$$f(T_g x,\lambda)=T_g f(x,\lambda)\quad\forall g\in\mathcal{G}, \tag{1.3}$$

where it is presumed henceforth that $T_g(\Omega)\subseteq\Omega$ for all $g\in\mathcal{G}$.

It is well known that (1.3) can be used to simplify what is often an intractable analysis of local bifurcation (cf. Sattinger [1979], Vanderbauwhede [1982], Golubitsky and Schaeffer [1985]). The purpose of this paper is to demonstrate that equivariance can also be exploited to considerable advantage in global bifurcation problems. The general approach is presented in § 2, where it is shown that a reduced problem for (1.1) can be constructed in a straightforward manner. The analysis of a reduced problem via well-known continuation theorems (cf. Rabinowitz [1973], Alexander and Yorke [1976]) then shows that symmetry properties are preserved on global continua of solutions. The procedure is illustrated in § 3 in the analysis of a nonlinearly elastic ring with $O(2)$ symmetry, and yields several new results. Symmetries of global post-buckling solution branches are enumerated, which enables a detailed qualitative analysis. In particular, it is shown that all global (primary) bifurcating branches with distinct symmetries are mutually nonintersecting.

* Received by the editors December 8, 1986; accepted for publication (in revised form) June 18, 1987. This research was supported in part by National Science Foundation grant DMS-8519918 and Air Force Office of Scientific Research grant AFOSR-86-0185.

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

2. Reduction and continuation. Let $\mathscr{H} \subseteq \mathscr{G}$ be a subgroup (not necessarily proper) and define

$$B_{\mathscr{H}} \equiv \{u \in B: T_g u = u \ \forall g \in \mathscr{H}\}, \tag{2.1}$$

which is called the *$\mathscr{H}$-fixed-point set.* It is straightforward to show that $B_{\mathscr{H}} \subseteq B$ is a linear subspace. Suppose that $B_{\mathscr{H}}$ is complemented in B, viz.,

$$B = B_{\mathscr{H}} \oplus A_{\mathscr{H}}. \tag{2.2}$$

It can be shown that (2.2) holds if and only if there exists a continuous projection $P_{\mathscr{H}}: B \to B$ with $\mathscr{R}(P_{\mathscr{H}}) = B_{\mathscr{H}}$ (cf. Rudin [1973, § 5.16]). If $\mathscr{H}$ is a compact group, then $P_{\mathscr{H}}$ is given explicitly by

$$P_{\mathscr{H}} x \equiv \int_{\mathscr{H}} T_g x \, d\mu(g) \quad \forall x \in B, \tag{2.3}$$

where μ is the Haar measure of $\mathscr{H}$. In the sequel, the decomposition (2.2) is assumed to hold.

By virtue of (1.3) and (2.1), it follows that

$$T_g f(u, \lambda) = f(u, \lambda) \quad \forall (u, \lambda) \in \Omega_{\mathscr{H}}, \quad g \in \mathscr{H}. \tag{2.4}$$

Thus, $f(u, \lambda) \in B_{\mathscr{H}}$ for all $(u, \lambda) \in \Omega_{\mathscr{H}}$ and $f: \Omega_{\mathscr{H}} \to B_{\mathscr{H}}$, where $\Omega_{\mathscr{H}} \equiv \Omega \cap (B_{\mathscr{H}} \times \mathbb{R})$. This leads to the following conclusion.

THEOREM 2.1. *Define $f_{\mathscr{H}} \equiv P_{\mathscr{H}} \circ f | \Omega_{\mathscr{H}}$. A point $(u_0, \lambda_0) \in \Omega_{\mathscr{H}}$ is a solution of* (1.1) *if and only if it is a solution of the $\mathscr{H}$-reduced problem*

$$f_{\mathscr{H}}(u, \lambda) = 0. \tag{2.5}$$

The solution set of (2.5), *denoted by $\Sigma_{\mathscr{H}}$, is called the $\mathscr{H}$-solution set.*

Henceforth, if $B_{\mathscr{H}}$ is finite-dimensional, then all such $(m \geqq 1)$-times continuously differentiable maps $f_{\mathscr{H}}: \Omega_{\mathscr{H}} \to B_{\mathscr{H}}$ are considered. If $B_{\mathscr{H}}$ is infinite-dimensional, then $f_{\mathscr{H}}$ is assumed to be of the form

$$f_{\mathscr{H}}(u, \lambda) \equiv u - c_{\mathscr{H}}(u, \lambda), \tag{2.6}$$

where $c_{\mathscr{H}}: \Omega_{\mathscr{H}} \to B_{\mathscr{H}}$ is completely continuous.

Recall that a *regular* solution point $(x_0, \lambda_0) \in \Sigma$ of (1.1) is one at which the Fréchet derivative $D_1 f^0 \equiv D_1 f(x_0, \lambda_0): B \to B$ is invertible. If $D_1 f^0$ is not invertible, then (x_0, λ_0) is called a *singular* point. Regular and singular points of the $\mathscr{H}$-reduced problem are defined analogously and are henceforth referred to as $\mathscr{H}$-regular and $\mathscr{H}$-singular points, respectively.

It often happens in applications that at least one solution point of (1.1), say, $(u_0, \lambda_0) \in \Sigma$, is known a priori. Suppose that $u_0 \in \Omega_{\mathscr{H}}$, for a particular subgroup $\mathscr{H} \subseteq \mathscr{G}$. Then $(u_0, \lambda_0) \in \Sigma_{\mathscr{H}}$ by Theorem 2.1. If (u_0, λ_0) is $\mathscr{H}$-regular, then the implicit function theorem guarantees the existence of a unique, local branch of solutions of (2.5) through (u_0, λ_0). By Theorem 2.1, that branch is a local, $\mathscr{H}$-symmetric solution path of (1.1). Let Σ^0 denote the connected component of Σ containing (u_0, λ_0). The following extension of the implicit function theorem shows that $\mathscr{H}$ symmetry is preserved globally.

EQUIVARIANT CONTINUATION THEOREM 2.2. *Let $(u_0, \lambda_0) \in \Sigma_{\mathscr{H}}$ be $\mathscr{H}$-regular. Then there exists a global, $\mathscr{H}$-symmetric solution branch of* (1.1) *through (u_0, λ_0). That is, there exists a connected subset $\Sigma^0_{\mathscr{H}} \subseteq \Sigma_{\mathscr{H}} \cap \Sigma^0$ containing (u_0, λ_0) that is characterized by at least one of the following alternatives:*

(i) *$\Sigma^0_{\mathscr{H}}$ is unbounded in $B \times \mathbb{R}$.*

(ii) *$\overline{\Sigma^0_{\mathscr{H}}} \cap \partial\Omega \neq \varnothing$.*

(iii) *$\Sigma^0_{\mathscr{H}} - \{(u_0, \lambda_0)\}$ is connected in $B \times \mathbb{R}$.*

Proof. In the context of the $\mathcal{H}$-reduced problem (2.5), this is precisely a theorem due to Alexander and Yorke [1976]. The claims hold for (1.1) by virtue of Theorem 2.1. □

Remark 2.1. This simple but profound result implies that the symmetry of an $\mathcal{H}$-regular point cannot completely "die out" somewhere along Σ^0. This has important applications in the construction of efficient numerical algorithms for global bifurcation problems (cf. Healey [1988a]). On the other hand, $\Sigma^0_{\mathcal{H}}$ need not coincide with Σ^0. Rather $\Sigma^0_{\mathcal{H}} \subseteq \Sigma^0$, and $\Sigma^0 \backslash \Sigma^0_{\mathcal{H}}$ (if it is nonempty) is the $\mathcal{H}$-symmetry-breaking component of Σ^0, which branches from $\Sigma^0_{\mathcal{H}}$. Indeed, an $\mathcal{H}$-regular point need not be a regular point of (1.1), i.e., (u_0, λ_0) may be an $\mathcal{H}$-symmetry-breaking bifurcation point. The $\mathcal{H}$ regularity of such a point can be exploited to enable its accurate computation in a numerical setting (cf. Werner and Spence [1984], Healey [1988a]).

Consider next the problem of bifurcation from a trivial branch of solutions of (1.1). Assume that

$$f(0, \lambda) = 0 \quad \forall (0, \lambda) \in \Omega. \tag{2.7}$$

The set $\Sigma^0_t \equiv (\{0\} \times \mathbb{R}) \cap \Omega \subset \Sigma_{\mathcal{G}}$ is called the trivial solution branch of (1.1), and is assumed to be homeomorphic to $\{0\} \times \mathbb{R}$. A solution pair $(0, \lambda_0) \in \Sigma^0_t$ is said to be a *bifurcation point* of (1.1) if every neighborhood of $(0, \lambda_0)$ contains solution pairs $(u_*, \lambda_*) \in \Sigma$ with $u_* \neq 0$. Let Σ^0 denote the connected component of Σ that contains $(0, \lambda_0)$. Define $L(\lambda) \equiv D_1 f(0, \lambda)$ and assume that $L(\lambda_0): B \to B$ is singular, which is necessary for $(0, \lambda_0)$ to be a bifurcation point.

It often happens in applications that some, but not all, of the $\mathcal{G}$ symmetry is broken on a bifurcating branch. Accordingly, suppose that $\mathcal{N}(L(\lambda_0)) \not\subset B_{\mathcal{G}}$. When $\mathcal{G}$ characterizes highly redundant symmetry (e.g., $\mathcal{G} = O(3)$), then $\dim \mathcal{N}(L(\lambda_0))$ is often quite large (cf. for example, Knightly and Sather [1980], Ihrig and Golubitsky [1984]). Thus, the analysis of local bifurcation at $(0, \lambda_0)$ can become intractable ($\dim \mathcal{N}(L(\lambda_0)) \geqq 3$ is enough (cf. Iooss and Joseph [1980])). This motivates seeking a reduced problem that simplifies the analysis of (1.1) near $(0, \lambda_0)$.

For any subgroup $\mathcal{H} \subseteq \mathcal{G}$, it follows from (2.5) and (2.7) that

$$f_{\mathcal{H}}(0, \lambda) = 0 \quad \forall (0, \lambda) \in \Omega_{\mathcal{H}}, \tag{2.8}$$

i.e., Σ^0_t is also the trivial solution branch of (2.5). The goal is to determine a subgroup (or subgroups) that yields a problem having nontrivial solutions in common with those of (1.1) near $(0, \lambda_0)$. A well-known procedure (in local bifurcation theory) for finding $\mathcal{H}$ is to seek a null vector, $y \in \mathcal{N}(L(\lambda_0))$, such that the *isotropy subgroup* of $\mathcal{G}$ at y,

$$\mathcal{H} \equiv \{g \in G \mid T_g y = y\}, \tag{2.9}$$

is proper. It then follows from Theorem 2.1 that $L_{\mathcal{H}}(\lambda_0) \equiv D_1 f_{\mathcal{H}}(0, \lambda_0) = [L(\lambda_0) \mid B_{\mathcal{H}}]: B_{\mathcal{H}} \to B_{\mathcal{H}}$ is singular. This suggests the following theorem.

EQUIVARIANT BIFURCATION THEOREM 2.3. *Suppose that f is C^2 and there exists a null vector $y \in \mathcal{N}(A(\lambda_0))$ that defines a proper isotropy subgroup $\mathcal{H}$ (cf.* (2.9)*). Assume the following*:

(EB1) $\dim \mathcal{N}(L_{\mathcal{H}}(\lambda_0))$ *is odd.*

(EB2) $[L'_{\mathcal{H}}(\lambda_0)]v \notin \mathcal{R}(L_{\mathcal{H}}(\lambda_0)) \quad \forall v \in \mathcal{N}(L_{\mathcal{H}}(\lambda_0))$.

Then $(0, \lambda_0)$ *is a bifurcation point of* (1.1) *such that in every sufficiently small neighborhood*

of $(0, \lambda_0)$, *there are nontrivial solutions* $(u_*, \lambda_*) \in \Sigma_{\mathscr{H}}$. *In particular, if* $\dim \mathscr{N}(L_{\mathscr{H}}(\lambda_0)) = 1$, *then there exists a unique, local, bifurcating branch of solutions of the form* $s \mapsto (\hat{u}(s), \hat{\lambda}(s)) \in \Sigma_{\mathscr{H}}$. *Moreover, there exists a connected subset* $\zeta^0_{\mathscr{H}} \subseteq \Sigma_{\mathscr{H}} \cap \Sigma^0 \backslash \Sigma^0_t$ *containing* $(0, \lambda_0)$ *that is characterized by at least one of the following properties*:

(i) $\zeta^0_{\mathscr{H}}$ *is unbounded in* $B \times \mathbb{R}$.

(ii) $\zeta^0_{\mathscr{H}} \cap \partial\Omega \neq \varnothing$.

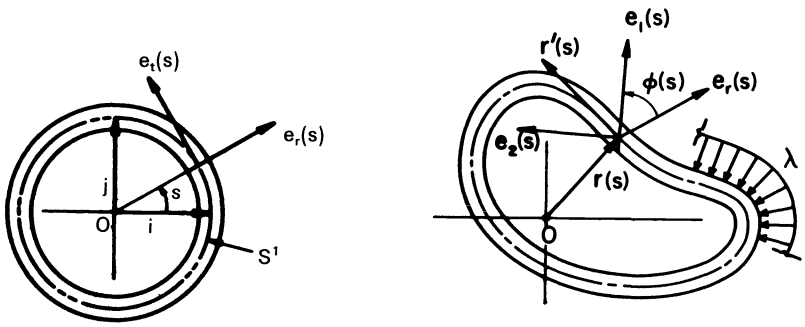
(iii) *There exists a pair* $(0, \lambda_*) \in \zeta^0_{\mathscr{H}} \cap \Sigma^0_t$ *with* $\lambda_* \neq \lambda_0$.

The subset $\zeta^0_{\mathscr{H}}$ *is called a global,* $\mathscr{H}$*-symmetric, bifurcating branch of* (1.1) *through* $(0, \lambda_0)$.

Proof. In the context of the $\mathscr{H}$-reduced problem (2.5), the first claim is Krasnosel'skii's theorem [1965], the second statement holds by a theorem of Crandall and Rabinowitz [1971], and the third result is due to Rabinowitz [1973]. Conditions (EB1) and (EB2) insure that the Leray-Schauder degree of $u \mapsto f_{\mathscr{H}}(u, \lambda)$ (which is well defined because of (3.1); cf., for example, Cronin [1964]) changes sign as λ passes through λ_0 along $u = 0$ (cf. Alexander and Fitzpatrick [1980]). That all claims hold for the full problem (1.1) is a direct consequence of Theorem 2.1. □

Remark 2.2. The claims of Theorem 2.3 pertaining to local bifurcation are well known and have appeared in various different forms (cf. Cicogna [1981], Vanderbauwhede [1982], Sattinger [1983], and Golubitsky [1983]). However, in those treatments the Lyapunov-Schmidt technique (cf., for example, Golubitsky and Schaeffer [1985]) is employed before the use of group-theoretic reasoning, thus obviating global conclusions.

3. Large post-buckling of a nonlinearly elastic ring with $O(2)$ symmetry. The analysis of a planar, nonlinearly elastic, circular ring under hydrostatic pressure (cf. Fig. 1) is presented in this section as an application of Theorem 2.3. The techniques of § 2 enable a detailed qualitative analysis of global post-buckling solution branches. The ring is modeled as a nonlinearly elastic rod that is capable of suffering geometrically exact stretching, shearing and bending. The formulation is due to Antman [1973]. In that work, global existence theorems were established, and some symmetry properties of solutions were obtained by phase-plane techniques. Similar results for an unshearable ring have been obtained by Antman [1970a], [1970b], where the latter work also includes a local bifurcation analysis from a radially symmetric state. A global bifurcation analysis of a circular arch (which leads to a two-point boundary-value problem involving the same differential equations as the ring in Antman [1973]) is presented in Antman and Dunn [1980].



undeformed configuration deformed configuration

FIG. 1. *Planar configurations of a ring.*

Let $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$ be a fixed orthonormal basis for E^3, Euclidean 3-space. The curve of centroids of the cross sections of the undeformed ring is taken to coincide with the unit circle $S^1 \subset \text{span}\{\mathbf{i}, \mathbf{j}\}$, along which points are identified by arclength $s \in \mathbb{R}_{2\pi} \equiv \mathbb{R}(\text{mod } 2\pi)$ (cf. Fig. 1). Define the unit vectors

$$\mathbf{e}_r(s) \equiv \cos(s)\mathbf{i} + \sin(s)\mathbf{j} \quad \text{and} \quad \mathbf{e}_t(s) \equiv -\sin(s)\mathbf{i} + \cos(s)\mathbf{j}. \tag{3.1}$$

Then $\{\mathbf{e}_r, \mathbf{e}_t, \mathbf{k}\}$ is an orthonormal frame field for E^3 on S^1.

A planar configuration of the ring is described by a mapping

$$\mathbb{R}_{2\pi} \ni s \mapsto (\mathbf{r}(s), \phi(s)) \in \text{span}\{\mathbf{i}, \mathbf{j}\} \times \mathbb{R}, \tag{3.2}$$

where $s \in \mathbb{R}_{2\pi}$ emphasizes the 2π-periodicity of $\mathbf{r}$ and ϕ. The vector $\mathbf{r}(s)$ is the position (measured from the center 0 of the ring) of the material point on the deformed centroidal curve that has position vector $\mathbf{e}_r(s)$ on the undeformed centroidal curve. The unit vector

$$\mathbf{e}_1[\phi(s), s] \equiv \cos[\phi(s)]\mathbf{e}_r(s) + \sin[\phi(s)]\mathbf{e}_t(s) \tag{3.3}$$

characterizes the deformed orientation of the cross section at s that has undeformed orientation coincident with $\mathbf{e}_r(s)$. Thus, $\phi(s)$ is a measure of rotation between the deformed and undeformed cross section at s. Defining

$$\mathbf{e}_2[\phi(s), s] \equiv -\sin[\phi(s)]\mathbf{e}_r(s) + \cos[\phi(s)]\mathbf{e}_t(s), \tag{3.4}$$

it follows that $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{k}\}$ is also an orthonormal frame field for E^3 on S^1.

The strain field of the rod is defined by

$$(\mathbf{r}', \phi') \quad \text{on } \mathbb{R}_{2\pi}, \tag{3.5}$$

where $(\,)'$ denotes differentiation with respect to argument on $\mathbb{R}_{2\pi}$. It is convenient to express $\mathbf{r}$ and $\mathbf{r}'$ as

$$\mathbf{r} = r_1\mathbf{e}_1 + r_2\mathbf{e}_2 \quad \text{and} \quad \mathbf{r}' = \xi_1\mathbf{e}_1 + \xi_2\mathbf{e}_2 \quad \text{on } \mathbb{R}_{2\pi}, \tag{3.6}$$

and to define

$$\xi_3 \equiv \phi' \quad \text{on } \mathbb{R}_{2\pi}. \tag{3.7}$$

Then by (3.1), (3.6), and (3.7), the strain-displacement relations for the ring are

$$r_1' - (1 + \xi_3)r_2 - \xi_1 = 0, \tag{3.8a}$$

$$r_2' + (1 + \xi_3)r_1 - \xi_2 = 0, \tag{3.8b}$$

$$\phi' - \xi_3 = 0 \qquad \text{on } \mathbb{R}_{2\pi}. \tag{3.8c}$$

The condition

$$\mathbf{r}' \cdot \mathbf{e}_2 = \xi_2 > 0 \quad \text{on } \mathbb{R}_{2\pi} \tag{3.9}$$

is imposed to insure that the local ratio of deformed length to undeformed length of centroidal curve does not vanish, and that $\mathbf{e}_1$ is not tangent to the deformed centroidal curve.

The undeformed ring is subjected to a uniform, hydrostatic pressure of intensity $\lambda > 0$ per unit (deformed) length of centroidal curve. Let $\mathbf{n}(s) \in \text{span}\{\mathbf{i}, \mathbf{j}\}$ denote the resultant force and $m(s)\mathbf{k}$ the resultant couple acting across the deformed cross section at s. The well-known equilibrium equations for the ring are

$$\mathbf{n}' + m'\mathbf{k} + \lambda\mathbf{k} \times \mathbf{r}' + \mathbf{r}' \times \mathbf{n} = \mathbf{0} \quad \text{on } \mathbb{R}_{2\pi}, \tag{3.10}$$

where "$\times$" denotes the usual right-handed cross product on E^3.

The ring is assumed to be homogeneous and nonlinearly elastic by requiring the existence of smooth constitutive functions

$$\mathbb{R}\times(0,\infty)\times\mathbb{R}\ni(\xi_1,\xi_2,\xi_3)\mapsto\hat{t}_i(\xi_1,\xi_2,\xi_3)\in\mathbb{R},\qquad i=1,2,3,$$

such that

$$\begin{aligned}\mathbf{n}(s)&=\sum_{i=1}^{2}\hat{t}_i(\xi_1(s),\xi_2(s),\xi_3(s))\mathbf{e}_i[\phi(s),s],\\ m(s)&=\hat{t}_3(\xi_1(s),\xi_2(s),\xi_3(s)).\end{aligned}\tag{3.11}$$

The constitutive functions are assumed to satisfy the following physically reasonable conditions (cf. Antman and Dunn [1980]):

The 3×3 Jacobian matrix

$$\left[\frac{\partial\hat{t}_i}{\partial\xi_j}\right]\text{ is positive-definite on }\mathbb{R}\times(0,\infty)\times\mathbb{R},\tag{3.12}$$

$$\hat{t}_1(0,\xi_2,\xi_3)=0\quad\forall\xi_2\in(0,\infty),\quad\xi_3\in\mathbb{R},\tag{3.13a}$$

$$\hat{t}_2(\xi_1,1,\xi_3)=0\quad\forall\xi_1,\quad\xi_3\in\mathbb{R},\tag{3.13b}$$

$$\hat{t}_3(\xi_1,\xi_2,0)=0\quad\forall\xi_1\in\mathbb{R},\quad\xi_2\in(0,\infty),\tag{3.13c}$$

$$\frac{\partial\hat{t}_2}{\partial\xi_1}(0,\xi_2,0)=\frac{\partial\hat{t}_2}{\partial\xi_3}(0,\xi_2,0)=0\quad\forall\xi_2\in(0,\infty),\tag{3.14}$$

$$\lim_{\xi_2\searrow0}\hat{t}_2(0,\xi_2,0)=-\infty.\tag{3.15}$$

Substitution of (3.11) into (3.10) leads to the following componential form of the equilibrium equations with respect to $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{k}\}$:

$$[\hat{t}_1(\xi_1,\xi_2,\xi_3)]'-(1+\xi_3)\hat{t}_2(\xi_1,\xi_2,\xi_3)-\lambda\xi_2=0,\tag{3.16a}$$

$$[\hat{t}_2(\xi_1,\xi_2,\xi_3)]'+(1+\xi_3)\hat{t}_1(\xi_1,\xi_2,\xi_3)+\lambda\xi_1=0,\tag{3.16b}$$

$$[\hat{t}_3(\xi_1,\xi_2,\xi_3)]'+\xi_1\hat{t}_2(\xi_1,\xi_2,\xi_3)-\xi_2\hat{t}_1(\xi_1,\xi_2,\xi_3)=0\quad\text{on }\mathbb{R}_{2\pi}.\tag{3.16c}$$

The systems (3.8) and (3.16) together with the constraint (3.9) constitute the field equations for the ring, where r_1, r_2, ϕ, ξ_i, $i=1,2,3$, are each 2π-periodic.

It is convenient to express the governing equations in a more compact form. Define the six-tuple

$$\mathcal{y}\equiv(r_1,r_2,\phi,\xi_1,\xi_2,\xi_3)\in\mathbb{R}^6.\tag{3.17}$$

Let $C^0_{2\pi}$ denote the Banach space of all continuous, 2π-periodic functions from $\mathbb{R}_{2\pi}$ into $\mathbb{R}^6$ with norm

$$|\mathcal{y}|=\max_{s\in\mathbb{R}_{2\pi}}\|\mathcal{y}(s)\|,$$

where $\|\ell\|$ denotes the Euclidean norm of $\ell\in\mathbb{R}^6$. Let $C^1_{2\pi}\subset C^0_{2\pi}$ denote the Banach space of all such mappings that are also continuously differentiable with norm

$$|\mathcal{y}|_1=|\mathcal{y}|+|\mathcal{y}'|.$$

Define the subset

$$\mathcal{O}\equiv\{\mathcal{y}\in C^1_{2\pi}:\xi_2>0\text{ on }\mathbb{R}_{2\pi}\}\times(0,\infty).\tag{3.18}$$

Consider a mapping $\mathcal{k}: \mathcal{O} \to C^0_{2\pi}$ defined by identifying the real-valued component functions h_1, h_2, and h_3 with the left sides of (3.16a), (3.16b), and (3.16c), respectively, and h_4, h_5, and h_6 with the left sides of (3.8a), (3.8b), and (3.8c), respectively. Then the field equations are equivalent to

$$\mathcal{k}(\mathcal{y}, \lambda) = 0. \tag{3.19}$$

Since the ring is homogeneous, it follows that the proper orthogonal group $SO(2)$ is a symmetry group of the ring. However, the material properties of the ring are assumed to concur with the complete geometric symmetry group $O(2)$. That is, the ring posesses through-thickness properties that are taken to be invariant under reflections as well as rotations in the plane. This places the following further restrictions on the constitutive functions:

$$\begin{aligned} &\hat{t}_1(-\xi_1, \xi_2, \xi_3) = -\hat{t}_1(\xi_1, \xi_2, \xi_3), \\ &\hat{t}_i(-\xi_1, \xi_2, \xi_3) = \hat{t}_i(\xi_1, \xi_2, \xi_3), \qquad i = 2, 3, \\ &\qquad\qquad \forall (\xi_1, \xi_2, \xi_3) \in \mathbb{R} \times (0, \infty) \times \mathbb{R}. \end{aligned} \tag{3.20}$$

Remark 3.1. The restrictions (3.20) can be justified by viewing the rod as a constrained, two-dimensional, homogeneous, isotropic elastic body, with the constitutive functions $\hat{t}_i$, $i = 1, 2, 3$, defined by appropriate resultants over the cross section (cf., for example, Antman and Carbone [1977, § 8]).

The appropriate representation of $O(2)$ is defined as follows. Let $T_g: C^m_{2\pi} \to C^m_{2\pi}$ $(m = 0, 1)$ be given by

$$[T_g \mathcal{y}](s) \equiv \begin{cases} \mathcal{y}(s+g) & \forall g \in \mathbb{R}_{2\pi} \cong SO(2), \\ E\mathcal{y}(g-s) & \forall g \in \mathbb{R}_{2\pi} \cong O(2) \backslash SO(2), \end{cases} \tag{3.21}$$

where $E: \mathbb{R}^6 \to \mathbb{R}^6$ is defined by $E\mathcal{y} \equiv (r_1, -r_2, -\phi, -\xi_1, \xi_2, \xi_3)$. It is straightforward to show that $g \mapsto T_g$ is a representation of $O(2)$ on $C^m_{2\pi}$. By (3.18) and (3.21), it is readily demonstrated that $T_g(\mathcal{O}) \subseteq \mathcal{O}$ for all $g \in \mathcal{G}$.

THEOREM 3.1. *The mapping in* (3.19) *is equivariant under T, i.e.,*

$$h(T_g \mathcal{y}, \lambda) = T_g h(\mathcal{y}, \lambda) \quad \forall g \in O(2).$$

Proof. Since the system (3.8), (3.16) is autonomous, the claim is immediate for all $g \in SO(2)$. If $g \in O(2) \backslash SO(2)$, it follows from (3.21) that

$$[T_g \mathcal{y}](s) = (r_1, -r_2, -\phi, -\xi_1, \xi_2, \xi_3)(g - s),$$

and

$$\frac{d}{ds}[T_g \mathcal{y}](s) = (-r_1', r_2', \phi', \xi_1', -\xi_2', -\xi_3')(g - s).$$

Substitution of these expressions into (3.8) and (3.16), while making use of (3.20), then yields the desired result. □

Intuitively, it is natural to seek a "trivial" radially symmetric solution (the existence of which is well known). Nonetheless, it is of interest to demonstrate that Theorem 2.1 leads to the same result by constructing the $O(2)$-reduced problem. By (2.3) and (3.21), the projection onto the $O(2)$-fixed-point set is given by

$$P\mathcal{y} = \frac{1}{2}\left[\frac{1}{2\pi}\int_0^{2\pi} \mathcal{y}(s+g)\, dg + \frac{1}{2\pi}\int_0^{2\pi} E\mathcal{y}(g-s)\, dg\right]. \tag{3.22}$$

By (3.17), (3.21), and the 2π-periodicity of y, it follows that

$$\frac{1}{2\pi}\int_0^{2\pi} y(s+g)\,dg = (\bar{r}_1, \bar{r}_2, \bar{\phi}, \bar{\xi}_1, \bar{\xi}_2, \bar{\xi}_3) \in \mathbb{R}_6,$$

and

$$\frac{1}{2\pi}\int_0^{2\pi} Ey(g-s)\,dg = (\bar{r}_1, -\bar{r}_2, -\bar{\phi}, -\bar{\xi}_1, \bar{\xi}_2, \bar{\xi}_3) \in \mathbb{R}_6,$$

where $\bar{r}_1$ denotes the average value of r_1 on $\mathbb{R}_{2\pi}$, etc. Thus, (3.22) becomes

$$Py = (\bar{r}_1, 0, 0, 0, \bar{\xi}_2, \bar{\xi}_3) \quad \text{on } \mathbb{R}_{2\pi}. \tag{3.23}$$

Note that (3.23) includes a reduction in both independent and dependent variables. By Theorem 2.1, the substitution of (3.23) into (3.19) identically satisfies (3.16b), (3.16c), and (3.8a), which is easily verified, and (3.8c) reduces to $\bar{\xi}_3 = 0$. The remaining nonzero algebraic equations are

$$\hat{t}_2(0, \bar{\xi}_2, 0) + \lambda\bar{\xi}_2 = 0, \qquad \bar{r}_1 = \bar{\xi}_2. \tag{3.24}$$

By (3.12) and (3.15), it follows that (3.24) has a unique solution $\bar{r}_1 = \bar{\xi}_2 = \psi(\lambda)$ for all $\lambda \geqq 0$, where $\psi : [0, \infty) \to (0, 1]$ is monotonically decreasing.

To investigate bifurcation from the $O(2)$-symmetric solution, it is convenient to bring (3.19) into the form of (2.7). Define

$$\bar{y}(\lambda) = (\psi(\lambda), 0, 0, 0, \psi(\lambda), 0) \quad \text{on } \mathbb{R}_{2\pi}, \tag{3.25}$$

and let $y = \bar{y}(\lambda) + x$, where

$$x = (x_1, x_2, x_3, x_4, x_5, x_6) \in C^1_{2\pi}. \tag{3.26}$$

Then set

$$f(x, \lambda) \equiv h(\bar{y}(\lambda) + x, \lambda) = 0. \tag{3.27}$$

In view of (3.18), $f : \Omega \to C^1_{2\pi}$, where

$$\Omega \equiv \{x \in C^1_{2\pi} : x_5 > -\psi(\lambda) \text{ on } \mathbb{R}_{2\pi}\} \times (0, \infty). \tag{3.28}$$

Clearly, (3.27) possesses the trivial solution $\{0\} \times (0, \infty)$. By (3.21) and (3.25), $T_g\bar{y}(\lambda) = \bar{y}(\lambda)$ for all $g \in O(2)$. Then by Theorem 3.1, the mapping in (3.27) is also equivariant

$$f(T_g x, \lambda) = T_g f(x, \lambda) \quad \forall g \in O(2). \tag{3.29}$$

The linearization of (3.27) about the trivial solution $x = 0$ is given by

$$L(\lambda)x = 0, \tag{3.30}$$

where $L(\lambda) \equiv D_1 f(0, \lambda) = D_1 h(\bar{y}(\lambda), \lambda) : C^1_{2\pi} \to C^0_{2\pi}$. By (3.8), (3.13), (3.14), (3.16), (3.19), and (3.24), the system (3.30) is given explicitly by

$$\hat{t}^0_{1,1}(\lambda)x_4' - [\hat{t}^0_{2,2}(\lambda) + \lambda]x_5 + \lambda\psi(\lambda)x_6 = 0, \tag{3.31a}$$

$$\hat{t}^0_{2,2}(\lambda)x_5' + [\hat{t}^0_{1,1}(\lambda) + \lambda]x_4 = 0, \tag{3.31b}$$

$$\hat{t}^0_{3,3}(\lambda)x_6' - \psi(\lambda)[\hat{t}^0_{1,1}(\lambda) + \lambda]x_4 = 0, \tag{3.31c}$$

$$x_1' - x_2 - x_4 = 0, \tag{3.31d}$$

$$x_2' + x_1 - x_5 + \psi(\lambda)x_6 = 0, \tag{3.31e}$$

$$x_3' - x_6 = 0 \quad \text{on } \mathbb{R}_{2\pi}, \tag{3.31f}$$

where

$$\hat{t}^0_{i,i}(\lambda) \equiv \frac{\partial \hat{t}_i}{\partial \xi_i}(0, \psi(\lambda), 0), \qquad i = 1, 2, 3.$$

A straightforward computation shows that (3.31) has three linearly independent null solutions for *any* value of λ, viz.,

$$\begin{aligned} &\Phi_1(s) \equiv (\cos(s), -\sin(s), 0, 0, 0, 0), \\ &\Phi_2(s) \equiv (\sin(s), \cos(s), 0, 0, 0, 0), \\ &\Phi_3(s) \equiv (0, 0, 1, 0, 0, 0,). \end{aligned} \tag{3.32}$$

Moreover, it is easy to demonstrate that whenever (x_0, λ_0) is a solution pair of (3.27), then so is $(x_0 + \sum_{i=1}^{3} c_i \Phi_i, \lambda_0)$, for arbitrary constants c_i, $i = 1, 2, 3$. This is quite reasonable from a physical point of view, since the ring is free to translate and rotate rigidly in the plane spanned by **i** and **j**. It can be shown that any rigid displacement of the ring has the form $\sum_{i=1}^{3} c_i \Phi_i$.

The system (3.31) also possesses nontrivial solution pairs $(x, \lambda) = (v_1^n, \lambda_n), (v_2^n, \lambda_n)$ whenever λ_n satisfies the characteristic equation

$$q(\lambda) = \left[1 + \frac{\lambda}{\hat{t}^0_{1,1}(\lambda)}\right]\left[1 + \frac{\lambda}{\hat{t}^0_{2,2}(\lambda)} + \frac{\lambda[\psi(\lambda)]^2}{\hat{t}^0_{3,3}(\lambda)}\right] = n^2, \tag{3.33}$$

where $n \geqq 2$ is an integer. By (3.12), $q(0) = 1$ (cf. Remark 3.2) and q is a strictly positive function on $\lambda > 0$. As discussed by Antman and Dunn [1980], (3.33) may have no roots, one root, or many roots, for a given integer n, depending on the behavior of the function q. Henceforth, it is assumed that q is monotonically increasing on the interval $(0, a]$, where $2 < a < \infty$. Thus, (3.33) has at least one root λ_n for each integer $n \in [2, \sqrt{q(a)}]$. There are two linearly independent null vectors associated with each root λ_n:

$$\begin{aligned} v_1^n(s) \equiv (&\alpha_1^n \cos(ns), \alpha_2^n \sin(ns), \alpha_3^n \sin(ns), \alpha_4^n \sin(ns), \alpha_5^n \cos(ns), \\ &n\alpha_3^n \cos(ns)), \\ v_2^n(s) \equiv (&-\alpha_1^n \sin(ns), \alpha_2^n \cos(ns), \alpha_3^n \cos(ns), \alpha_4^n \cos(ns), -\alpha_5^n \sin(ns), \\ &-n\alpha_3^n \sin(ns)), \end{aligned} \tag{3.34}$$

where

$$\begin{aligned} &\alpha_1^n = -\frac{1}{n}(\alpha_2^n + \alpha_4^n), \\ &\alpha_2^n = \frac{1}{n^2 - 1}[-n^2 \psi(\lambda_n)\alpha_3^n + \alpha_4^n + n\alpha_5^n], \\ &\alpha_3^n = \frac{-\psi(\lambda_n)}{n^2 \hat{t}^0_{3,3}(\lambda_n)}[\hat{t}^0_{1,1}(\lambda_n) + \lambda_n]\alpha_4^n, \\ &\alpha_4^n \neq 0, \\ &\alpha_5^n = \frac{[\hat{t}^0_{1,1}(\lambda_n) + \lambda_n]}{n\hat{t}^0_{2,2}(\lambda_n)}\alpha_4^n. \end{aligned}$$

Remark 3.2. It is interesting to note that if λ_1 is a root of (3.33) for $n = 1$ ($\lambda = 0$ is always such a root), then the 2π-periodicity conditions dictate that (3.31) evaluated

at $\lambda = \lambda_1$ admits only the trivial solution and the rigid-body solutions (3.32). This contrasts with results from several previous works on circular rings where strain formulations were employed without accounting for displacements (cf. Antman [1973], Tadjbakhsh and Odeh [1967]). Such an analysis is equivalent to considering only (3.31a), (3.31b), and (3.31c), which admit nontrivial solutions at $\lambda = \lambda_1$ similar to (3.34) for (x_4, x_5, x_6). This in turn, ostensibly implies the existence of bifurcating solutions (x_4, x_5, x_6) of least period 2π to the full nonlinear problem. In the above-cited works, subtle techniques were required to demonstrate the inadmissibility of such solutions. In this formulation, the existence of such solutions (bifurcating from the trivial solution) never arises.

The analysis of bifurcation is complicated for two reasons. The linearized operator in (3.30) has a three-dimensional kernel for all $\lambda \in (0, \infty)$, and it admits two additional null vectors for all values of λ satisfying (3.33). The latter condition is common in problems with $O(2)$ symmetry (cf., for example, Vanderbauwhede [1982]). Following (2.9), consider the null vector v_1^n and note that

$$v_1^n(s + 2\pi j/n) = v_1^n(s),$$
$$Ev_1^n(2\pi j/n - s) = v_1^n(s), \qquad j = 1, 2, \cdots, n.$$

By (3.21), it follows that the isotropy subgroup of $O(2)$ at v_1^n is

$$D_n \cong \{g \in O(2): T_g v_1^n = v_1^n\},$$

where D_n denotes the dihedral group of order $2n$. The "mode shapes" associated with v_1^n for $n = 2, 3, 4, 5$ are shown schematically in Fig. 2. This motivates constructing the D_n-reduced problem.

Remark 3.3. It is easy to show that any linear combination of the modes $v_1^n(s)$ and $v_2^n(s)$ (suitably scaled) is equivalent to a change of phase; $v_1^n(s+g) = [T_g v_1^n](s)$

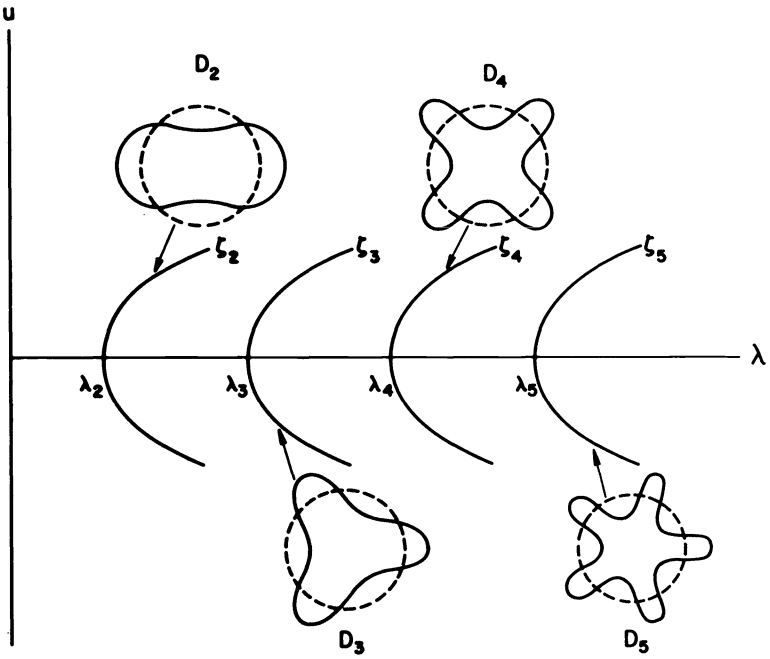


FIG. 2. *Schematic bifurcation diagram for* $\mu_{2n} > 0$, $n = 2, 3, 4, 5$.

for some $g \in SO(2)$. The isotropy subgroup of $O(2)$ at $T_g v_1^n$ is $gD_ng^{-1} \equiv \{ghg^{-1}: h \in D_n\}$, which is conjugate to D_n. It is shown later in this section that the solutions of a gD_ng^{-1}-reduced problem can be generated from the solutions of the D_n-reduced problem.

By Theorem 2.1, the D_n-reduced problem for (3.27) consists of the same mapping $\mathscr{f}$ restricted to the D_n-fixed-point set. By (2.3), the projection onto that subspace is given by

$$P_n x = \frac{1}{2n} \sum_{j=1}^{n} [x(s + 2\pi j/n) + Ex(2\pi j/n - s)]. \tag{3.35}$$

THEOREM 3.2. *Let $C^m_{2\pi/n}$ denote the space of all m-times continuously differentiable functions from $\mathbb{R}_{2\pi}$ into $\mathbb{R}^6$ that are $2\pi/n$-periodic. Let $B^m_{2\pi/n}$ denote the D_n-fixed-point set of $C^m_{2\pi}$. Then $B^m_{2\pi/n}$ is the subspace of all $u = (u_1, u_2, u_3, u_4, u_5, u_6) \in C^m_{2\pi/n}$ such that the component functions $u_i(s)$ are even for $i = 1, 5, 6$ and odd for $i = 2, 3, 4$, i.e.,*

$$B^m_{2\pi/n} = \{u \in C^m_{2\pi/n}: u_i(-s) = u_i(s),\ u_j(-s) = -u_j(s),\ i = 1, 5, 6,\ j = 2, 3, 4\}. \tag{3.36}$$

Proof. For $x \in C^m_{2\pi/n}$, consider the ith component function of $P_n x(s) \equiv (u_1(s), u_2(s), \cdots, u_6(s)) \in B^m_{2\pi/n}$. By (3.21) and (3.35),

$$u_i(s) = \begin{cases} \dfrac{1}{2n} \displaystyle\sum_{j=1}^{n} [x_i(s + 2\pi j/n) + x_i(2\pi j/n - s)] & \text{for } i = 1, 5, 6, \\ \dfrac{1}{2n} \displaystyle\sum_{j=1}^{n} [x_i(s + 2\pi j/n) - x_i(2\pi j/n - s)] & \text{for } i = 2, 3, 4. \end{cases}$$

Clearly, $u_i(-s) = u_i(s)$ for $i = 1, 5, 6$, and $u_i(-s) = -u_i(s)$, $i = 2, 3, 4$. Moreover, the $2\pi/n$-periodicity of u_i follows from the 2π-periodicity of x_j, $j = 1, 2, \cdots, 6$, on $\mathbb{R}_{2\pi}$.

On the other hand, if $u \in B^m_{2\pi/n}$ as defined by (3.36), then it is straightforward to verify that

$$u(s + 2\pi j/n) = u(s),$$

and

$$Eu(2\pi j/n - s) = u(s), \qquad j = 1, 2, \cdots, n,$$

i.e., $T_g u = u$ for all $g \in D_n$, by virtue of (3.21). □

The D_n-reduced problem for (3.27) is given by

$$\mathscr{f}_n(u, \lambda) \equiv h(\bar{y}(\lambda) + u, \lambda) = 0, \tag{3.37}$$

where $\mathscr{f}_n: \Omega_n \to B^0_{2\pi/n}$, and where

$$\Omega_n \equiv \{u \in B^1_{2\pi/n}: u_5 > -\psi(\lambda) \text{ on } \mathbb{R}_{2\pi}\} \times (0, \infty). \tag{3.38}$$

The linearization of (3.37) about $u = 0$, denoted

$$L_n(\lambda) u = 0, \tag{3.39}$$

where $L_n(\lambda) \equiv D_1 \mathscr{f}_n(0, \lambda): B^1_{2\pi/n} \to B^0_{2\pi/n}$, is again defined by (3.31). However, the $2\pi/n$-periodicity of $u \in B^1_{2\pi/n}$, $n \geqq 2$, shows that Φ_1 and Φ_2 (cf. (3.32)) are not solutions of (3.39). Also, Φ_3 is not a solution, since u_3 is odd. Finally, the evenness and oddness of the various components of u imply that the only nontrivial solution of (3.39) at a root λ_n of (3.33) is v_1^n. That is, $\dim \mathscr{N}(L_n(\lambda_n)) = 1$, and condition (EB1) of Theorem 2.3 is satisfied.

Define the formal adjoint operator $L_n^*(\lambda)$ of $L_n(\lambda)$ with respect to the inner product

$$\langle y, x \rangle \equiv \frac{n}{2\pi} \int_0^{2\pi/n} y(s)^t x(s)\, ds, \tag{3.40}$$

where $\mathscr{y}^t x \equiv \sum_{i=1}^n y_i x_i$. A straightforward computation shows that $L_n^*(\lambda)\mathscr{y} = 0$ is given explicitly by

$$
\begin{aligned}
&y_4' - y_5 = 0, \qquad y_5' + y_4 = 0, \qquad y_6' = 0,\\
&\hat{t}^0_{1,1}(\lambda) y_1' - [\hat{t}^0_{1,1}(\lambda) + \lambda] y_2 + \psi(\lambda)[\hat{t}^0_{1,1}(\lambda) + \lambda] y_3 + y_4 = 0,\\
&\hat{t}^0_{2,2}(\lambda) y_2' + [\hat{t}^0_{2,2}(\lambda) + \lambda] y_1 + y_5 = 0,\\
&\hat{t}^0_{3,3}(\lambda) y_3' - \lambda\psi(\lambda) y_1 - \psi(\lambda) y_5 + y_6 = 0 \quad \text{on } \mathbb{R}_{2\pi/n}.
\end{aligned} \tag{3.41}
$$

For each root λ_n, $n \geqq 2$, of (3.33), the system (3.41) admits nontrivial solutions

$$
\begin{aligned}
v_1^{n^*}(s) &\equiv (\beta_1^n \cos(ns), -\beta_2^n \sin(ns), \beta_3^n \sin(ns), 0, 0, 0),\\
v_2^{n^*}(s) &\equiv (\beta_1^n \sin(ns), \beta_2^n \cos(ns), -\beta_3^n \cos(ns), 0, 0, 0),
\end{aligned} \tag{3.42}
$$

where

$$
\begin{aligned}
&\beta_1^n \neq 0,\\
&\beta_2^n = \frac{1}{n}\left[1 + \frac{\lambda_n}{\hat{t}^0_{2,2}(\lambda_n)}\right]\beta_1^n,\\
&\beta_3^n = \frac{\lambda_n \psi(\lambda_n)}{n\hat{t}^0_{3,3}(\lambda_n)}\beta_1^n,
\end{aligned}
$$

such that $\langle v_i^{n^*}, v_j^n \rangle = \delta_{ij}$. However, it is easily demonstrated that $\langle v_2^{n^*}, u \rangle = 0$ for all $u \in B^0_{2\pi/n} \Rightarrow v_2^{n^*} \notin B^{0^*}_{2\pi/n}$, the dual space of $B^0_{2\pi/n}$. In particular, $v_2^{n^*} \notin \mathcal{N}(L_n^*(\lambda_n)) \subset B^{0^*}_{2\pi/n}$. Thus, $L_n^*(\lambda_n)$ has only $v_1^{n^*}(s)$ as a null vector, and $L_n(\lambda_n)$ has a simple zero eigenvalue. By the alternative theorem (cf., for example, Stakgold [1979]), the transversality condition (EB2) of Theorem 2.3 then reduces to

$$
\langle v_1^{n^*}, [L_n'(\lambda_n)] v_1^n \rangle \neq 0. \tag{3.43}
$$

A lengthy but straightforward computation employing (3.13), (3.24), (3.31), and (3.33) yields

$$
\begin{aligned}
\langle v_1^{n^*}, [L_n'(\lambda_n)] v_1^n \rangle = -K_n \Bigg\{ &1 + \frac{\lambda_n}{\hat{t}^0_{2,2}(\lambda_n)} + n^2 \frac{\hat{t}^0_{1,1}(\lambda_n)[\hat{t}^0_{2,2}(\lambda_n) + \lambda_n]}{[\hat{t}^0_{1,1}(\lambda_n) + \lambda_n]^2}\\
&+ \frac{\lambda_n \psi(\lambda_n)}{[\hat{t}^0_{2,2}(\lambda_n)]^2}\hat{t}^0_{2,22}(\lambda_n) + \frac{[\psi(\lambda_n)]^2}{\hat{t}^0_{3,3}(\lambda_n)}[\hat{t}^0_{2,2}(\lambda_n) - \lambda_n]\Bigg\},
\end{aligned} \tag{3.44}
$$

where $K_n > 0$ is a constant and $\hat{t}^0_{2,22}(\lambda) \equiv (\partial^2 \hat{t}_2 / \partial \xi_2^2)(0, \psi(\lambda), 0)$. By (3.12), it follows that the first three terms inside the brackets on the right side of (3.44) are strictly positive. However, the last two terms can each be positive, zero, or negative, depending upon the behavior of the function $\xi_2 \mapsto \hat{t}_2(0, \xi_2, 0)$. There are no (known) natural hypotheses for the second derivative $\hat{t}^0_{2,22}(\lambda)$. Thus, the right side of (3.44) could conceivably vanish at a given root λ_n. In such case, a more refined local analysis is required (cf. Rabier [1985]), which is beyond the scope of this work. At any rate, if (3.43) is satisfied, then Theorem 2.3 leads to the following result.

THEOREM 3.3. *If* (3.43) *holds, then there exists a unique, local, nontrivial,* D_n*-symmetric branch of solutions of* (3.27) *of the form* $(x, \lambda) = (\hat{u}_n(t), \hat{\lambda}_n(t)) \in \Omega_n$ *for all* $|t| < \varepsilon$ *with* $(\hat{u}_n(0), \hat{\lambda}_n(0)) = (0, \lambda_n)$. *Moreover,* $\hat{u}_n(t) = tv_1^n + o(t)$ *as* $t \to 0$.

If $\mathscr{f}_n$ is sufficiently smooth, then a local analysis of (3.37) at $(0, \lambda_n)$ shows that the bifurcation diagram is a "pitchfork," i.e., $\hat{\lambda}_n(t) = \lambda_n + \mu_{2n}t^2 + o(t)$ as $t \to 0$. That is, the coefficient of the first-order term vanishes, viz., $\mu_{1n} \equiv \langle v_1^{n^*}, D_1^2\mathscr{f}_n(0, \lambda_n)[v_1^n, v_1^n]\rangle = 0$, where $D_1^2\mathscr{f}_n(0, \lambda_n): B^1_{2\pi/n} \times B^1_{2\pi/n} \to B_{2\pi/n}$ is the second (Fréchet) derivative of $\mathscr{f}_n$ with respect to its first argument evaluated at $(u, \lambda) = (0, \lambda_n)$. To see this, note that (3.21), (3.34), and (3.42) lead to

$$
\begin{aligned}
&[T_{\pi/n}v_1^n](s) = -v_1^n(s),\\
&[T_{\pi/n}v_1^{n^*}](s) = -v_1^{n^*}(s) \quad \text{for } \frac{\pi}{n} \in SO(2).
\end{aligned} \tag{3.45}
$$

Successive differentiation of (3.29) with respect to x leads to

$$
D_1^2\mathscr{f}(0, \lambda_n)[T_g v, T_g w] = T_g D_1^2\mathscr{f}(0, \lambda_n)[v, w], \tag{3.46}
$$

for all $g \in O(2)$, $v, w \in C^1_{2\pi/n}$. Evaluating (3.46) at $g = \pi/n \in SO(2)$ and $v = w = v_1^n$, and then taking its inner product with $T_{\pi/n}v_1^{n^*}$ yields

$$
\begin{aligned}
\langle T_{\pi/n}v_1^{n^*}, D_1^2\mathscr{f}(0, \lambda_n)[T_{\pi/n}v_1^{n^*}, T_{\pi/n}v_1^n]\rangle &= \langle T_{\pi/n}v_1^{n^*}, T_{\pi/n}D_1^2\mathscr{f}(0, \lambda_n)[v_1^n, v_1^n]\rangle\\
&= \langle v_1^{n^*}, D_1^2\mathscr{f}(0, \lambda_n)[v_1^n, v_1^n]\rangle,
\end{aligned} \tag{3.47}
$$

where $(3.47)_2$ follows from (3.21), (3.40), and (3.42). By virtue of Theorem 2.1, it follows that $D_1^2\mathscr{f}_n(0, \lambda_n)[v, w] \equiv D_1^2\mathscr{f}(0, \lambda_n)[v, w]$ for all $v, w \in B^1_{2\pi/n}$. Hence, (3.45) and (3.47) lead to

$$
-\langle v_1^{n^*}, D_1^2\mathscr{f}_n(0, \lambda_n)[v_1^n, v_1^n]\rangle = \langle v_1^{n^*}, D_1^2\mathscr{f}_n(0, \lambda_n)[v_1^n, v_1^n]\rangle \Rightarrow \mu_{1n} = 0.
$$

It can be shown that μ_{2n} is the quotient of $-\langle v_1^{n^*}, D_1^3\mathscr{f}_n(0, \lambda_n)[v_1^n, v_1^n, v_1^n]\rangle$ and the right side of (3.44), (cf., for example, Golubitsky and Schaeffer [1985]). A schematic bifurcation diagram for $\mu_{2n} > 0$, $n = 2, 3, 4, 5$, is presented in Fig. 2.

To make the conclusions of Theorem 3.3 global via Theorem 2.3, it is sufficient to demonstrate that (3.37) can be recast into the form (2.6). Referring back to (3.8), (3.16), (3.19), and (3.27), it follows that (3.37) can be expressed as

$$
[M_n(u, \lambda)]u' - \mathscr{h}_n(u, \lambda) = 0, \tag{3.48}
$$

where $\mathscr{h}_n: \mathbb{R}^6 \times (0, \infty) \to \mathbb{R}^6$ is smooth and M_n is a smooth, (6×6)-matrix-valued function on $\mathbb{R}^6 \times (0, \infty)$. M_n is given explicitly by

$$
M(u, \lambda) = \left[\begin{array}{c|c} [0] & [\partial \hat{t}_i/\partial \xi_j](\bar{\mathscr{y}}(\lambda) + u) \\ \hline [I] & [0] \end{array}\right],
$$

where $[0]$ is the 3×3 zero matrix, $[I]$ is the 3×3 identity matrix and $[\partial \hat{t}_i/\partial \xi_j](\bar{\mathscr{y}}(\lambda) + u)$ is the 3×3 Jacobian of the constitutive functions (cf. (3.11)) evaluated at $(\xi_1, \xi_2, \xi_3) = (u_4, \psi(\lambda) + u_5, u_6)$. By (3.12), $M_n(u, \lambda)$ is invertible for all $(u, \lambda) \in \Omega_n$. Thus, (3.48) leads to

$$
u' = [M_n(u, \lambda)]^{-1}\mathscr{h}_n(u, \lambda), \tag{3.49}
$$

where

$$
[M_n(u, \lambda)]^{-1} = \left[\begin{array}{c|c} [0] & [I] \\ \hline [\partial \hat{t}_i/\partial \xi_j]^{-1}(\bar{\mathscr{y}}(\lambda) + u) & [0] \end{array}\right].
$$

Equation (3.49) is equivalent to the integral equation (cf. Krasnosel'skii and Zabreiko [1984, § 28])

$$u(s) = u(2\pi/n) + \int_0^s [M_n(u(\tau), \lambda)]^{-1} \not{p}_n(u(\tau), \lambda)\, d\tau. \tag{3.50}$$

The right side of (3.50) defines a mapping $c_n : \Omega_n \to B^1_{2\pi/n}$. Moreover, c_n is readily shown to be completely continuous by virtue of the continuity of the integrand and the Arzelà-Ascoli Theorem. Thus, (3.50) is of the form (2.6).

Let Σ_n denote the solution set of (3.37), and denote the trivial solution by $\Sigma_t = \{0\} \times (0, \infty)$. The following theorem is a consequence of the final part of Theorem 2.3.

THEOREM 3.4. *Let the hypothesis of Theorem* 3.3 *hold. Then the D_n-symmetric bifurcating branch of* (3.27) *from* $(0, \lambda_n)$ *is global. That is, there exists a connected set* $\zeta_n \subset \overline{\Sigma_n \backslash \Sigma_t} \subset \Omega_n$ *containing* $(0, \lambda_n)$ *that is characterized by at least one of the following properties*:

(i) *ζ_n is unbounded in* $C^1_{2\pi}(0, \infty)$.

(ii) $\zeta_n \cap \partial\Omega \neq \varnothing$.

(iii) *There exists a pair* $(0, \lambda_*) \in \zeta_n \cap \Sigma_t$ *with* $\lambda_* \neq \lambda_n$.

The D_n symmetry of each global branch ζ_n, $n \in N \cap [2, \sqrt{q(a)}]$, does not directly imply which properties (i)-(iii) of Theorem 3.4 actually characterize ζ_n. Indeed, by (2.1), $D_n \subset D_{pn} \Rightarrow \Sigma_{pn} \subseteq \Sigma_n$ for all $n, p \in N$, $n \geqq 2$. In particular, this does not preclude the possibility that $\zeta_{pn} \cap \zeta_n \neq \varnothing$. However, a further analysis of the reduced problem (3.50) or equivalently (3.37) shows that this cannot occur.

For any $u \in B^1_{2\pi/n}$, $n \geqq 2$, it follows from the continuity, oddness, and $2\pi/n$-periodicity of the component function u_4 (which represents the shear strain) that $u_4(j\pi/n) = 0, j = 1, 2, \cdots, 2n$. Let $S^1_{2\pi/n}$ denote the open set of all $u \in B^1_{2\pi/n}$ such that u_4 has exactly $2n$ simple zeros at $s = j\pi/n, j = 1, 2, \cdots, 2n$, on $\mathbb{R}_{2\pi}$. An analysis similar to that of Rabinowitz [1973, Lemma 2.7], based on the fact that $v^1_n \in S^1_{2\pi/n}$, shows that $u \in S^1_{2\pi/n}$ for all $(u, \lambda) \in \mathcal{N}_n \cap \zeta_n$, where $\mathcal{N}_n \subset \Omega_n$ is some sufficiently small neighborhood of $(0, \lambda_n)$. That is, u_4 inherits the nodal properties of the corresponding eigenfunction locally along ζ_n.

Now the nodal behavior of u_4 can change only if u is in the closure of $S^1_{2\pi/n}$ somewhere along $\zeta_n \Rightarrow \exists$ at least one number $\tau \in \mathbb{R}_{2\pi}$ such that $u_4(\tau) = u_4'(\tau) = 0$, i.e., u_4 has a double zero at $s = \tau$ (cf. Rabinowitz [1973]). Consider the degenerate initial-value problem (3.37) subject to $u_4(\tau) = u_4'(\tau) = 0$. By (3.19), (3.25), and (3.27), this is equivalent to finding a solution of (3.8) and (3.16) of the form $(r_1, r_2, \phi, \xi_1, \xi_2, \xi_3) = (\psi(\lambda) + u_1, u_2, u_3, u_4, \psi(\lambda) + u_5, u_6)$, where $u \in \Omega_n$, subject to the same degenerate initial conditions. The evaluation of (3.16) and (3.8) at $s = \tau$, while making use of (3.13a), leads to

$$(1 + u_6(\tau))\tilde{t}_2(\lambda, \tau) + \lambda[\psi(\lambda) + u_5(\tau)] = 0, \tag{3.16a)''}$$

$$\tilde{t}_{2,2}(\lambda, \tau)u_5'(\tau) + \tilde{t}_{2,3}(\lambda, \tau)u_6'(\tau) = 0, \tag{3.16b)''}$$

$$\tilde{t}_{3,2}(\lambda, \tau)u_5'(\tau) + \tilde{t}_{3,3}(\lambda, \tau)u_6'(\tau) = 0, \tag{3.16c)''}$$

$$u_1'(\tau) = [1 + u_6(\tau)]u_2(\tau), \tag{3.8a)''}$$

$$u_2'(\tau) = -[1 + u_6(\tau)]u_1(\tau) + u_5(\tau) - \psi(\lambda)u_6(\tau), \tag{3.8b)''}$$

$$u_3'(\tau) = u_6(\tau), \tag{3.8c)''}$$

where

$$\tilde{f}_2(\lambda, \tau) \equiv \hat{f}_2(0, \psi(\lambda) + u_5(\tau), u_6(\tau)),$$

$$\tilde{f}_{i,j}(\lambda, \tau) \equiv \frac{\partial \hat{f}_i}{\partial \xi_j}(0, \psi(\lambda) + u_5(\tau), u_6(\tau)).$$

From (3.12), it follows that (3.16b)″ and (3.16c)″ have the unique solution $u_5'(\tau) = u_6'(\tau) = 0$. By (3.8c)″, $u_6(\tau) = \prime \Rightarrow u_3^0(\tau) = 0$, in which case (3,16a)″ has the unique solution $u_5(\tau) = 0$, by virtue of (3.24). Thus $(r_1, r_2, \xi_1, \xi_2, \xi_3) = (\psi(\lambda), 0, u_3(\tau), 0, \psi(\lambda), 0)$ is a critical point of the system (3.8) and (3.16)$\Rightarrow u = (0, 0, u_3(\tau), 0, 0, 0)$ is a critical point of (3.37). Moreover, $u \in B^1_{2\pi/n} \Rightarrow u_3$ is odd $\Rightarrow u_3(\tau) = 0$. Thus, $u \equiv 0$ is the unique solution of (3.37) subject to the degenerate initial conditions. However, this implies that u_4 can change its nodal properties along ζ_n only at the trivial solution branch, i.e., branches with distinct symmetries do not intersect. This leads to the following strengthened version of Theorem 3.4.

THEOREM 3.5. *Let the hypothesis of Theorem* 3.3 *hold. Then in addition to the claims of Theorem* 3.4, *it follows that* $\zeta_n - \{(0, \lambda_n)\} \subset S^1_{2\pi/n} \times (0, \infty)$. *In particular,* $S^1_{2\pi/n} \cap S^1_{2\pi/m} = \varnothing$ *for* $m \neq n \Rightarrow \zeta_n \cap \zeta_m = \varnothing$. *Thus,* ζ_n *is characterized by property* (iii) *of Theorem* 3.4 *only if* λ_* *and* λ_n *are two distinct roots of* (3.33) *corresponding to the same integer* $n \geqq 2$.

A particular case of interest is when $q: (0, \infty) \to (0, \infty)$ (cf. (3.33)) is monotonically increasing with $q \to \infty$ as $\lambda \to \infty$, in which case (3.33) has exactly one root λ_n for each integer $n \geqq 2$. Then by Theorem 3.5, ζ_n is either unbounded or $\zeta_n \cap \partial\Omega \neq \varnothing$. By (3.28), the latter condition is characterized by the existence of a solution point $(u^0, \lambda^0) \in \zeta_n$ such that $\xi_2^0 \equiv u_5^0 + \psi(\lambda) = 0$ somewhere on $\mathbb{R}_{2\pi}$ and/or $\lambda^0 = 0$. With appropriate growth conditions on the constitutive functions (for which there is little or no physical basis), the possibility of violating the unilateral constraint (3.9) can be eliminated by the existence theorems of Antman [1973].

To conclude the analysis of the ring, note that if (x_0, λ_0) is a solution point of (3.27), then so is $(T_g x_0, \lambda_0)$ for all $g \in O(2)$, by virtue of (3.29). In particular, if $(0, \lambda_n) \neq (u_0, \lambda_0) \in \zeta_n$, then $\{(T_g u_0, \lambda_0): 0 \leqq g < 2\pi/n,\ g \in SO(2)\}$ is a one-parameter family or an orbit of *distinct* solutions of (3.27). By (3.21), $[T_g u_0](s) = u_0(s + g)$ for all $g \in SO(2)$, which has the physical interpretation of a clockwise rotation of the buckled shape corresponding to $u_0 \in \zeta_n$ (cf. Fig. 2) through the angle g. That is, u_0 and $T_g u_0$ differ by only a change of phase. Now $T_g u_0$ is an element of the $(gD_n g^{-1})$-fixed-point set (cf. Remark 3.3), i.e., $T_h[T_g u_0] = T_g u_0$ for all $h \in gD_n g^{-1}$. Consequently,

$$\Gamma_n \equiv \{(T_g u, \lambda): (u, \lambda) \in \zeta_n, 0 \leqq g < 2\pi/n, g \in SO(2)\}$$

is a global, connected "sheet" of bifurcating solutions from $(0, \lambda_n)$, with each "slice" $T_g\zeta_n \equiv \{(T_g u, \lambda): (u, \lambda) \in \zeta_n\} \subset \Gamma_n$ being a global bifurcating branch of solutions of the $(gD_n g^{-1})$-reduced problem. Finally, it follows from (3.21) and (3.28) that $|u|_1 = |T_g u|_1$ and $(u, \lambda) \in \partial\Omega \Rightarrow (T_g u, \lambda) \in \partial\Omega$, for all $g \in O(2)$, $u \in C^1_{2\pi/n}$. Thus, if ζ_n is unbounded and/or $\zeta_n \cap \partial\Omega \neq \varnothing$, then $T_g\zeta_n$ is unbounded and/or $T_g\zeta_n \cap \partial\Omega \neq \varnothing$, respectively.

4. Concluding remarks. The analysis of the ring problem is noteworthy for several reasons. First, many results are new and complement the general existence theorems of Antman [1973]. In that work symmetry properties of solutions that agree with those obtained here were deduced by clever phase-plane arguments. However, the results of this work are sharper. Symmetries of specific global solution branches are enumerated, and it is shown that primary bifurcating branches with distinct symmetries are mutually disjoint. Further, in contrast to phase-plane methods, the systematic techniques of § 2 are applicable to a broad class of equivariant operators.

The ring problem demonstrates other advantages in analyzing a reduced problem that are not apparent in the abstract development of § 2. The D_n-reduced problem is designed to eliminate the troublesome analysis of bifurcation that is associated with a two-dimensional null space in the presence of $O(2)$ symmetry (cf. (3.34)). However, the D_n-reduced problem also eliminates the rigid-body solutions (cf. (3.32)), thus yielding a standard bifurcation problem with a one-dimensional kernel.

Of more importance is the fact that the D_n-reduced problem admits a detailed qualitative analysis. The technique of identifying nodal properties of a solution along global bifurcating branches was devised by Crandall and Rabinowitz [1970] for nonlinear Sturm–Liouville problems. Since that time, there have been numerous applications of the method, mostly due to Antman and his co-workers, to more general two-point boundary-value problems. The application of such techniques to the full ring equations and to problems with $O(2)$ symmetry, in general, is not clear. However, the D_n-reduced problem fixes the indeterminate phase and admits only solutions with shear strains (in particular) that are odd with period $2\pi/n$. Consequently, it was possible to show that the nodal properties of the shear strain on $\mathbb{R}_{2\pi}$ are inherited from the corresponding eigenfunction of the linearized problem and are preserved on global bifurcating branches. That the shear strain plays a crucial role is hardly surprising in view of the work of Antman and Dunn [1980] on circular arches.

It is interesting to note that if the symmetry group of a ring is $SO(2)$, but not $O(2)$, then the subgroups are the cyclic groups C_n. It can be shown that the state variable of the C_n-reduced problem is an element of $C^1_{2\pi/n}$. In particular, the various component functions do not possess the evenness and oddness that play a crucial role in the detailed qualitative analysis of an $O(2)$-symmetric ring.

Finally, it should be pointed out that the construction of a reduced problem presented in § 2 also holds for multi-parameter problems. Moreover, it can be easily extended to more general classes of operator equations (cf. Healey [1988b]). Thus, it is clear that the analysis of reduced problems by other global continuation theorems, e.g., Alexander and Antman [1981], [1983] and Fitzpatrick and Pejsachowicz [1986], leads to obvious generalizations of Theorems 2.2 and 2.3.

REFERENCES

J. C. ALEXANDER AND S. S. ANTMAN [1981], *Global and local behavior of bifurcating multidimensional continua of solutions for multiparameter nonlinear eigenvalue problems*, Arch. Rational Mech. Anal., 76, pp. 339–354.

——— [1983], *Global behavior of solutions of nonlinear equations depending on infinite-dimensional parameters*, Indiana Univ. Math. J., 32, pp. 39–62.

J. C. ALEXANDER AND P. M. FITZPATRICK [1979], *The homotopy of certain spaces of nonlinear operators, and its relation to global bifurcation of the fixed points of parameterized condensing operators*, J. Funct. Anal., 34, pp. 87–106.

——— [1980], *Galerkin approximations in several parameter bifurcation problems*, Math. Proc. Soc., 87, pp. 489–500.

J. C. ALEXANDER AND J. A. YORKE [1976], *The implicit function theorem and global methods of cohomology*, J. Funct. Anal., 21, pp. 330–339.

S. S. ANTMAN [1970a], *Existence of solutions of the equilibrium equations for nonlinearly elastic rings and arches*, Indiana Univ. Math. J., 20, pp. 281–302.

——— [1970b], *The shape of buckled nonlinearly elastic rings*, Z. Angew. Math. Phys., 21, pp. 422–438.

——— [1973], *Monotonicity and invertibility conditions in one-dimensional nonlinear elasticity*, in Nonlinear Elasticity, R. Dickey, ed., Academic Press, New York, pp. 57–92.

S. S. ANTMAN AND E. CARBONE [1977], *Shear and necking instabilities in nonlinear elasticity*, J. Elasticity, 7, pp. 127–151.

S. S. ANTMAN AND E. DUNN [1980], *Qualitative behavior of buckled nonlinearly elastic arches*, J. Elasticity, 10, pp. 225-239.

G. CICOGNA [1981], *Symmetry breakdown from bifurcations*, Lett. Nuovo Cimento, 31, pp. 600-602.

M. CRANDALL AND P. RABINOWITZ [1970], *Nonlinear Sturm-Liouville eigenvalue problems and topological degree*, J. Math. Mech., 19, pp. 1083-1102.

——— [1971], *Bifurcation from simple eigenvalues*, J. Funct. Analysis, 8, pp. 321-340.

J. CRONIN [1964], *Fixed Points and Topological Degree in Nonlinear Analysis*, American Mathematical Society, Providence, RI.

P. M. FITZPATRICK AND J. PEJSACHOWICZ [1986], *An extension of the Leray-Schauder degree for fully nonlinear elliptic problems*, in Proc. Symposium on Pure Math., Vol. 45.

M. GOLUBITSKY [1983], *The Bénard problem, symmetry and the lattice of isotropy subgroups*, in Bifurcation Theory, Mechanics and Physics, C. P. Borter et al., eds., Reidel, Dordrecht, pp. 225-256.

M. GOLUBITSKY AND D. SCHAEFFER [1985], *Singularities and Groups in Bifurcation Theory*, Vol. I, Springer-Verlag, New York.

T. J. HEALEY [1988a], *A group-theoretic approach to computational bifurcation problems with symmetry*, Comput. Methods Appl. Mech. Engrg., to appear.

——— [1988b], *Symmetry and equivariance in nonlinear elastostatics* I: *differential field equations*, Arch. Rational Mech. Anal., to appear.

E. IHRIG AND M. GOLUBITSKY [1984], *Pattern selection with $O(3)$ symmetry*, Phys. D, 13, pp. 1-33.

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

G. KNIGHTLY AND D. SATHER [1980], *Buckled states of a spherical shell under uniform external pressure*, Arch. Rational Mech. Anal., 72, pp. 315-380.

M. A. KRASNOSEL'SKII [1965], *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Oxford.

M. A. KRASNOSEL'SKII AND P. P. ZABREIKO [1984], *Geometrical Methods of Nonlinear Analysis*, Springer-Verlag, New York.

P. RABIER [1985], *Topics in One-Parameter Bifurcation Problems*, Springer-Verlag, New York, Berlin.

P. RABINOWITZ [1973], *Some aspects of nonlinear eigenvalue problems*, Rocky Mountain J. Math., 3, pp. 161-202.

A. ROBERT [1983], *Introduction to the Representation Theory of Compact and Locally Compact Groups*, Cambridge University Press, Cambridge.

W. RUDIN [1973], *Functional Analysis*, McGraw-Hill, New York.

D. H. SATTINGER [1979], *Group Theoretic Methods in Bifurcation Theory*, Springer-Verlag, New York.

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

I. STAKGOLD [1979], *Green's Functions and Boundary Value Problems*, John Wiley, New York.

I. TADJBAKHSH AND F. ODEH [1967], *Equilibrium states of elastic rings*, J. Math. Anal. Appl., 18, pp. 59-74.

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

B. WERNER AND A. SPENCE [1984], *The computation of symmetry-breaking bifurcation points*, SIAM J. Numer. Anal., 21, pp. 388-399.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

006

ASYMPTOTIC PROPERTIES FOR INHOMOGENEOUS ITERATIONS OF NONLINEAR OPERATORS*

TAKAO FUJIMOTO† AND ULRICH KRAUSE‡

Abstract. The theorems on weak and strong ergodicity for inhomogeneous products of nonnegative matrices are extended to inhomogeneous iterations of nonlinear positive operators on Euclidean space. In particular some concave version of the Coale-Lopez theorem is presented and applied to a density-dependent Leslie model. The results are obtained, via Hilbert's projective pseudometric, from general theorems on inhomogeneous iterations of operators mapping a metric space into itself.

Key words. discrete dynamical systems, weak and strong ergodicity, Hilbert's projective metric, nonlinear Leslie model

AMS(MOS) subject classifications. 47H05, 47H10, 54H20, 92A15

1. Introduction. Consider a discrete dynamical system given by an operator f mapping the state space into itself. For the dynamical behaviour of the system of particular interest are asymptotic properties of the iterates f^n if $n \to \infty$. If, however, the system itself changes in the course of time with f_t as "law of motion" at, say, time t, then one will become interested in asymptotic properties of the inhomogeneous iterations $f_n \cdot f_{n-1} \cdot \cdots \cdot f_1$ if $n \to \infty$. This kind of problem arises in mathematical biology and mathematical economics where, e.g., the principle governing the growth of a population or the choice of a technology itself depends on time (cf. [4], [7], [8], [13], [14]). For the case of linear operators in finite dimensions, inhomogeneous iterations become inhomogeneous products $A_nA_{n-1} \cdots A_1$ of matrices, the asymptotic properties of which have been investigated for a long time in the theory of Markov chains and the theory of nonnegative matrices, respectively (cf. [14]). Since absolute magnitudes tend thereby to grow exponentially, one considers relative magnitudes as exemplified by $x_n = A_nA_{n-1} \cdots A_1x / \|A_nA_{n-1} \cdots A_1x\|$, x being a starting vector and $\|\cdot\|$ some vector space norm. There are two major stability results here, one on so-called strong ergodicity, meaning convergence of x_n for arbitrary starting vectors to the same limit and the other one on so-called weak ergodicity, meaning convergence of $x_n - y_n$ to 0 for any two starting vectors x, y, and y_n starting with y. These results have important applications, in particular to population dynamics, where the second result is known also as the Coale-Lopez theorem (cf. [4], [8], [13], [14]).

As in other fields too, linearity is a strong idealization concerning applications and on behalf of the latter results on nonlinear operators, e.g., concave ones, are requested. We present in § 3 as the main results of this paper theorems on weak and strong ergodicity for positive nonlinear operators in finite dimensions that contain the well-known theorems on nonnegative matrices as special cases. In particular, a concave version of the Coale-Lopez theorem is presented. Section 4 provides some concrete classes of nonlinear examples for these results by developing a density- and time-dependent version of the Leslie model of population dynamics.

To prove the theorems of § 3 we translate the order-theoretic framework of positive operators into a metric framework and prove in § 2 the corresponding theorems. The metric used is Hilbert's projective pseudometric which was first introduced by

* Received by the editors September 22, 1986; accepted for publication (in revised form) August 4, 1987.
† University of Kagawa, Takamatsu, Kagawa 760, Japan.
‡ University of Bremen, 2800 Bremen 33, Federal Republic of Germany.

Birkhoff [1], [2] into functional analysis. He used it to extend Jentzsch's theorem on linear integral operators to abstract linear operators by applying Banach's contraction mapping principle with respect to this metric (cf. also [10], [11]). Hilbert's metric is applied to inhomogeneous products of matrices in [8] and in [14, 2nd ed.] (cf. also [4]). In extending Birkhoff's theorem to nonlinear operators, Hilbert's metric is used in [11] within an infinite-dimensional nonlinear context (cf. also [12] for the finite-dimensional case). The present paper applies Hilbert's metric to inhomogeneous iterations of nonlinear operators in finite dimensions. (A direct approach to inhomogeneous iterations in finite dimensions that does not involve Hilbert's metric, but that is by no means more simple, can be found for matrices in [14, 1st ed.] and for nonlinear operators in [6].) In the case of homogeneous iterations usually one of the many variations of the contraction mapping principle is applied with respect to Hilbert's metric. To treat inhomogeneous iterations, however, something different is needed. Although there is an enormous literature on the contraction mapping principle, as surveyed, e.g., in [3], [5], and [10], there seems to be none handling the composition of several different contractions. Hence in § 2 we give a systematic account of inhomogeneous iterations within the metric framework. Although in the present paper the material of § 2 is used in § 3 only for the finite-dimensional case, it may also be applied to infinite dimensions.

2. Inhomogeneous iterations of operators mapping a metric space into itself. An operator $f: X \to X$ on a metric space X with metric d is said to be *nonexpansive* if $d(f(x), f(y)) \leqq d(x, y)$ for all $x, y \in X$; it is said *contractive on* Y, for $Y \subseteq X$, if

$$d(f(x), f(y)) < d(x, y) \quad \text{for all } x, y \in Y \text{ with } x \neq y.$$

We call a sequence $(f_n)_n$ of operators $f_n: X \to X$ an (asymptotically) *contractive sequence on* Y for $Y \subseteq X$, if there exists a continuous mapping $c: Y \times Y \to \mathbb{R}$ such that the following two conditions are satisfied:

(i) $c(x, y) < d(x, y)$ for all $x, y \in Y$ with $x \neq y$;

(ii) To every $\varepsilon > 0$ there exists a $N(\varepsilon) \in \mathbb{N}$ such that $d(f_n(x), f_n(y)) \leqq c(x, y) + \varepsilon$ for all $n \geqq N(\varepsilon)$, all $x, y \in Y$.

For a given $r \geqq 1$ and a given sequence $(f_n)_n$ of operators on X we will consider also the sequence of *lumped operators* $(F_m)_m$ defined by

$$F_m = f_{m+r-1} \cdot \cdots \cdot f_{m+1} \cdot f_m$$

(where $\cdot$ stands for the composition of mappings).

In what follows we are concerned with the asymptotic behaviour of inhomogeneous iterations, which means the behaviour of $f_n \cdot \cdots \cdot f_2 \cdot f_1(x)$ for $n \to \infty$ where $x \in X$ and $(f_n)_n$ is a sequence of operators on X. In the special case of (homogeneous) iteration the underlying sequence is simply $(f, f, \cdots)$ for some operator f on X. This is a contractive sequence precisely when f is a contractive operator, and the sequence of lumped operators in this case is $(f^r, f^r, \cdots)$.

The following theorem provides conditions under which inhomogeneous iterations come close together irrespective of the starting point. This does not necessarily mean that the iterations itself do converge. (In the next section we will see that the former is related to so-called weak ergodicity and the latter to strong ergodicity.)

THEOREM 1. *Let $(f_n)_n$ be a sequence of nonexpansive operators on the metric space (X, d) such that for some $r \geqq 1$ the sequence $(F_m)_m$ of lumped operators is contractive on Y and satisfies $F_m(X) \subseteq Y$ for some compact subset Y of X and almost all m. Then $\lim_{n\to\infty} d(x_n, y_n) = 0$ for any two sequences defined by $x_{n+1} = f_n(x_n)$ and $y_{n+1} = f_n(y_n)$ with arbitrary starting points $x_1, y_1 \in X$.*

Proof. (1) Consider first a sequence $(g_m)_m$ of nonexpansive operators, contractive on Y and satisfying $g_m(X) \subseteq Y$ for some compact $Y \subseteq X$ and almost all m. Let $x_{m+1} = g_m(x_m)$, $y_{m+1} = g_m(y_m)$ for $m \in \mathbb{N}$ and $x_1, y_1 \in X$ arbitrary. Since eventually $(x_m, y_m) \in Y \times Y$ and Y is compact, there exists a subsequence $(x_{k(m)}, y_{k(m)})_m$ converging to some $(x^*, y^*) \in Y \times Y$. By the nonexpansiveness of g_m

$$d(x_{m+1}, y_{m+1}) = d(g_m(x_m), g_m(y_m)) \leqq d(x_m, y_m),$$

and hence $\lim_{m\to\infty} d(x_m, y_m) = d(x^*, y^*)$.

The sequence $(g_{k(m)})_m$ is also contractive and according to the definition there exists a function c and to every $\varepsilon > 0$ an $M(\varepsilon)$ such that

$$d(x_{k(m)+1}, y_{k(m)+1}) = d(g_{k(m)}(x_{k(m)}), g_{k(m)}(y_{k(m)})) \leqq c(x_{k(m)}, y_{k(m)}) + \varepsilon$$

for all $m \geqq M(\varepsilon)$. Letting $m \to \infty$ from this we obtain $d(x^*, y^*) \leqq c(x^*, y^*) + \varepsilon$ because of $d(x^*, y^*) \leqq d(x_m, y_m)$ and the continuity of c. Since $\varepsilon > 0$ was arbitrary $d(x^*, y^*) \leqq c(x^*, y^*)$ which together with $c(x, y) < d(x, y)$ for $x \neq y$ and $x, y \in Y$ yields $x^* = y^*$. Thus finally $\lim_{m\to\infty} d(x_m, y_m) = 0$.

(2) Suppose now $(f_n)_n$ is a sequence as in the theorem and put $g_m = F_{(m-1)r+1}$. Being a composition of nonexpansive mappings, g_m is nonexpansive. Step (1) therefore yields $\lim_{m\to\infty} d(\bar{x}_m, \bar{y}_m) = 0$ for sequences defined by $\bar{x}_{m+1} = g_m(\bar{x}_m)$, $\bar{y}_{m+1} = g_m(\bar{y}_m)$, $\bar{x}_1 = x_1$, $\bar{y}_1 = y_1$. By the definition of lumped operators

$$g_m \cdot g_{m-1} \cdot \cdots \cdot g_2 \cdot g_1 = f_{mr} \cdot f_{mr-1} \cdot \cdots \cdot f_2 \cdot f_1,$$

and hence $\bar{x}_{m+1} = x_{mr+1}$, $\bar{y}_{m+1} = y_{mr+1}$. For any natural number n there exist nonnegative integers $m(n)$ and i such that $n = m(n)r + i$ with $0 \leqq i < r$. Since $x_{n+1} = f_n \cdot f_{n-1} \cdot \cdots \cdot f_{n-i+1}(x_{n-i+1})$ and the f_n's are nonexpansive it follows that

$$d(x_{n+1}, y_{n+1}) \leqq d(x_{m(n)r+1}, y_{m(n)r+1}) = d(\bar{x}_{m(n)+1}, \bar{y}_{m(n)+1}).$$

Thus $\lim_{n\to\infty} d(x_n, y_n) = 0$. □

Remark. The above proof shows that Theorem 1's conclusion remains valid if instead of the F_m's the operators $g_m = F_{(m-1)r+1}$ are considered with $g_m(X) \subseteq Y$ for some compact $Y \subseteq X$ and almost all m and such that the sequence $(g_m)_m$ contains a contractive subsequence on Y.

Now we are looking for conditions ensuring the convergence of the inhomogeneous iterations itself to a common limit for arbitrary starting points. Obviously this is stronger than the convergence statement made in Theorem 1 and therefore we have to add some assumptions.

THEOREM 2. *Let $(f_n)_n$ and $(F_m)_m$ be sequences of operators as in Theorem 1. Suppose in addition for the lumped operators F_m uniform convergence on the metric space to some operator F. This assumption is particularly fulfilled in case the operators f_n converge uniformly to some f. Then for arbitrary starting points $x_1 \in X$ the sequence defined by $x_{n+1} = f_n(x_n)$ converges to the unique fixed point of F, or f, respectively.*

Proof. (1) Consider first a sequence $(g_m)_m$ of nonexpansive operators, contractive on Y and satisfying $g_m(X) \subseteq Y$ for some compact $Y \subseteq X$ and almost all m. Assume g_m converges uniformly on (X, d) to some operator g, i.e., to $\varepsilon > 0$ there exists $N_1(\varepsilon)$ such that $d(g_m(x), g(x)) \leqq \varepsilon$ for all $m \geqq N_1(\varepsilon)$ and all $x \in X$. Since $(g_m)_m$ is contractive on Y there exists a function c and to $\varepsilon > 0$ there exists $N_2(\varepsilon)$ such that

$$d(g_m(x), g_m(y)) \leqq c(x, y) + \varepsilon \quad \text{for all } m \geqq N_2(\varepsilon) \text{ and all } x, y \in Y.$$

Therefore

$$\begin{aligned} d(g(x), g(y)) &\leqq d(g(x), g_m(x)) + d(g_m(x), g_m(y)) + d(g_m(y), g(y)) \\ &\leqq c(x, y) + 3\varepsilon \quad \text{for } m \geqq N_1(\varepsilon), m \geqq N_2(\varepsilon). \end{aligned}$$

This yields $d(g(x), g(y)) \leqq c(x, y) < d(x, y)$ for $x \neq y$, $x, y \in Y$, i.e., g is contractive on Y.

Let now $x_1 \in X$ be an arbitrary starting point, fixed in what follows. We want to show that the set A of all limit points of the set $\{x_m\}$, where $x_{m+1} = g_m(x_m)$, consists precisely of the unique fixed point of g. Since $(x_m)_m$ is eventually contained in the compact set Y, $A \subseteq Y$ and $A \neq \emptyset$. Pick some $x \in A$. Then there is a subsequence $(x_{j(n)})_n$ of $(x_n)_n$ converging to x and we may assume that $x_{j(n)-1} \in Y$ for all n. $(x_{j(n)-1})_n$ contains a subsequence $(x_{k(n)-1})_n$ converging to some $y \in Y$. Obviously

$$d(x, g(y)) \leqq d(x, x_{k(n)}) + d(g_{k(n)-1}(x_{k(n)-1}), g(x_{k(n)-1})) + d(g(x_{k(n)-1}), g(y))$$

and taking into account that $(x_{k(n)})_n$ converges to x, $(g_m)_m$ converges uniformly to g and that g is contractive on Y we obtain $d(x, g(y)) = 0$, i.e., $x = g(y)$. Because of $y \in A$ by the same argument we can find $y_2 \in A$ such that $y = g(y_2)$. By iteration we obtain to every $n \in \mathbb{N}$ a $y_n \in A$ such that $x = g^n(y_n)$. Since $y_n \in A \subseteq Y$, $(y_n)_n$ contains a subsequence $(y_{h(n)})_n$ converging to some $y^* \in Y$. Because of $g(Y) \subseteq Y$, g is contractive also on the metric space (Y, d) and hence by a well-known version of Banach's contraction mapping principle $(g^n(y^*))_n$ converges to the unique fixed point x^* of g. From $d(x, x^*) \leqq d(g^{h(n)}(y_{h(n)}), g^{h(n)}(y^*)) + d(g^{h(n)}(y^*, x^*)$ we therefore obtain $d(x, x^*) = 0$, i.e., $x = x^*$. This proves $A = \{x^*\}$, i.e., the convergence of the sequence $(x_m)_m$, defined by $x_{m+1} = g_m(x_m)$ with arbitrary starting point, to the unique fixed point of g.

(2) Let for arbitrary $x_1 \in X$ $(x_n)_n$ be defined by $x_{n+1} = f_n(x_n)$. We fix i with $0 \leqq i < r$ and define $g_m = F_{(m-1)r+i+1}$. By the assumptions made on the lumped operators F_m we may apply step (1) to the sequence $(g_m)_m$. Therefore the sequence defined by $\bar{x}_{m+1} = g_m(\bar{x}_m)$, $\bar{x}_1 = x_{i+1}$ converges to the unique fixed point x^* of $g = F$.

By the definition of lumped operators

$$g_m \cdot g_{m-1} \cdot \cdots \cdot g_1 = f_{mr+i} \cdot f_{mr+i-1} \cdot \cdots \cdot f_{i+1},$$

and hence $\bar{x}_{m+1} = x_{mr+i+1}$. Therefore for any fixed $0 \leqq i < r$ $(x_{mr+i+1})_m$ converges to the unique fixed point x^* of F. Since any natural n can be written as $n = mr + i$ with $0 \leqq i < r$ it follows that $(x_n)_n$ converges to x^*. This proves the theorem in case the lumped operators converge uniformly to some F.

(3) Let $(f_n)_n$ be an arbitrary sequence of operators on a metric space X converging uniformly on X to some uniformly continuous operator f. We first show that to every $\varepsilon > 0$ and to every natural k there exists an $N(\varepsilon, k)$ such that

$$(*) \qquad d(f^k(x), f_{n_1} \cdot f_{n_2} \cdot \cdots \cdot f_{n_k}(x)) \leqq \varepsilon \quad \text{for } n_i \geqq N(\varepsilon, k) \text{ and all } x \in X.$$

By assumption, for $\varepsilon > 0$ given there exist $\delta(\varepsilon) > 0$ and $N(\varepsilon)$ such that

$$d(f(x), f_n(y)) \leqq d(f(x), f(y)) + d(f(y), f_n(y)) \leqq \varepsilon/2 + \varepsilon/2 \leqq \varepsilon,$$

provided $d(x, y) \leqq \delta(\varepsilon)$ and $n \geqq N(\varepsilon)$.

Hence, $(*)$ holds for $k = 1$ with $N(\varepsilon, 1) = N(\varepsilon)$. Suppose, $(*)$ holds for some $k \geqq 1$. Then

$$d(f^k(x), f_{n_2} \cdot \cdots \cdot f_{n_{k+1}}(x)) \leqq \delta(\varepsilon) \quad \text{for all } n_2, \cdots, n_{k+1} \geqq N(\delta(\varepsilon), k).$$

Putting $N(\varepsilon, k+1) = \max\{N(\varepsilon), N(\delta(\varepsilon), k)\}$, we obtain

$$d(f^{k+1}(x), f_{n_1} \cdot f_{n_2} \cdot \cdots \cdot f_{n_{k+1}}(x)) = d(f(f^k(x)), f_{n_1}(f_{n_2} \cdot \cdots \cdot f_{n_{k+1}}(x))) \leqq \varepsilon$$

for $n_1, n_2, \cdots, n_{k+1} \geqq N(\varepsilon, k+1)$. This proves $(*)$ to hold for all k.

Now, let $(f_n)_n$ as in Theorem 2 and Theorem 1, respectively, and suppose uniform convergence to some operator f. Since every f_n is nonexpansive, f must be uniformly continuous. Thus (*) applies and yields for $\varepsilon > 0$ given $d(f^r(x), F_m(x)) \leqq \varepsilon$ for all $m \geqq N(\varepsilon)$ and all $x \in X$. Therefore $(F_m)_m$ converges uniformly on X to $F = f^r$. Step (2) then yields convergence of the sequence defined by $x_{n+1} = f_n(x_n)$, $x_i \in X$, to the unique fixed point x^* of F. By nonexpansiveness of f_n and uniform convergence to f, from

$$d(f(x^*), x^*) \leqq d(f(x^*), f(x_n)) + d(f(x_n), f_n(x_n)) + d(x_{n+1}, x^*)$$

it follows that $f(x^*) = x^*$. □

Remark. Theorem 2 remains valid, if instead of the whole sequence $(F_m)_m$ only some subsequence is required to be contractive on Y. The beginning of step (1) in the proof, when applied to this subsequence yields that F is contractive on Y. But then $(F_m)_m$ is contractive on Y too. For, it is true in general that a sequence of operators h_m converging uniformly to some contractive h must be contractive (with $c(x, y) = d(h(x), h(y))$).

From Theorem 2 we may derive the following criterion which is cast more directly in terms of the given operators.

THEOREM 3. *For any sequence $(f_n)_n$ of nonexpansive operators on the metric space X the sequence defined by $x_{n+1} = f_n(x_n)$ converges for arbitrary $x_1 \in X$ to the same limit point, provided the following condition is satisfied: For some $r \geqq 1$ the sequence $(F_m)_m$ of lumped operators converges uniformly on X to some operator F for which there exists an open and relatively compact subset U of X such that $F(X) \subseteq U$ and F is contractive on the closure $\bar{U}$.*

This condition is particularly satisfied if $(f_n)_n$ converges uniformly on X to some f for which $f^r(X) \subseteq U$ and f^r is contractive on $\bar{U}$ for some $r \geqq 1$, U being an open and relatively compact subset of X.

Proof. To derive the above criterion from Theorem 2 we only have to show that $(F_m)_m$ is contractive on Y and for almost all m $F_m(X) \subseteq Y$ for some compact subset Y of X. Putting $Y = \bar{U}$ F is contractive on Y and hence (cf. the remark following the proof of Theorem 2) the sequence $(F_m)_m$ is contractive on Y. It remains to show that $F_m(X) \subseteq Y$ for almost all m. To every $x \in X$ there exists some $\varepsilon(x) > 0$ such that $B(F(x), \varepsilon(x)) \subseteq U$, $B(F(x), \varepsilon(x))$ being the open ball with center $F(x)$ and radius $\varepsilon(x)$. Obviously the closure $\overline{F(X)}$ is contained in $\bigcup_{x \in X} B(F(x), \frac{1}{2}\varepsilon(x))$ and there is a finite cover, $\overline{F(X)} \subseteq \bigcup_{x \in \tilde{X}} B(F(x), \frac{1}{2}\varepsilon(x))$ for some finite set $\tilde{X} \subseteq X$, because $\overline{F(X)}$ is contained in the compact set $\bar{U}$. Let ε be the smallest of the numbers $\frac{1}{2}\varepsilon(x)$, $x \in \tilde{X}$. By uniform convergence of the lumped operators there exists $N(\varepsilon)$ such that $d(F_m(x), F(x)) \leqq \varepsilon$ for all $m \geqq N(\varepsilon)$ and all $x \in X$. For $x \in X$ there exists some $\tilde{x} \in \tilde{X}$ such that $F(x) \in B(F(\tilde{x}), \frac{1}{2}\varepsilon(\tilde{x}))$ and therefore

$$d(F_m(x), F(\tilde{x})) \leqq d(F_m(x), F(x)) + d(F(x), F(\tilde{x})) \leqq \tfrac{1}{2}\varepsilon(\tilde{x}) + \tfrac{1}{2}\varepsilon(\tilde{x}) = \varepsilon(\tilde{x})$$

for all $m \geqq N(\varepsilon)$. Hence $F_m(x) \in B(F(\tilde{x}), \varepsilon(\tilde{x})) \subseteq U$ for all $x \in X$, all $m \geqq N(\varepsilon)$. Thus $F_m(X) \subseteq Y$ for almost all m. □

Remark. In the situation of Theorem 3 the meaning of being contractive for the sequence of lumped operators is that some iterate of the limit function of the original sequence is contractive. More precisely, let $(f_n)_n$ be a sequence of nonexpansive operators converging uniformly on a metric space to some operator f. Then for any $r \geqq 1$, the sequence $(F_m)_m$ of lumped operators is contractive if and only if f^r is contractive. This is immediate by parts (1) and (3) in the proof of Theorem 2 and the remark thereafter.

3. Inhomogeneous iterations of nonlinear positive operators on Euclidean space. The results of the previous section we shall now apply to obtain the theorems of weak and strong ergodicity for inhomogeneous iterations of nonlinear positive operators on Euclidean space. As in the case of linear operators, these theorems have applications in mathematical biology. We shall obtain a concave version of the Coale-Lopez theorem of population dynamics as a corollary, which then is applied in the next section to a density-dependent Leslie model.

Let E denote the k-dimensional Euclidean space with typical element $x = (x_1, \cdot, \cdot, \cdot, x_k)$, $x_i \in \mathbb{R}$. For $x, y \in E$ we write $x \leqq y$ if $x_i \leqq y_i$ for all i; we write $x < y$ if $x_i < y_i$ for all i. E_+ denotes the positive cone $E_+ = \{x \in E \mid x \geqq 0\}$. By a *scale* we mean a continuous functional $p: E_+ \to \mathbb{R}_+$ with $p(x) = 0$ only for $x = 0$ and such that p is positively homogeneous, i.e., $p(\lambda x) = \lambda p(x)$ for $x \in E_+$, $\lambda \in \mathbb{R}_+$, and p is monotonic, i.e., $p(x) \leqq p(y)$ for $0 \leqq x \leqq y$. Obviously every monotonic norm is a scale, but there are others, e.g., maxima or minima of these norms. In what follows we fix on E_+ an arbitrary scale p and we denote its unit level set by X, $X = \{x \in E_+ \mid p(x) = 1\}$. This set X when equipped with a metric defined below will serve as the metric space underlying the previous section. As for the operators we shall employ various properties in the sequel. An operator $T: E_+ \to E_+$ is

proper if $Tx = 0$ is equivalent to $x = 0$;

subhomogeneous if for $x, y \in X$, $0 \leqq \lambda \leqq 1$, $\lambda x \leqq y$ implies $\lambda Tx \leqq Ty$;

ray-preserving if for every $x \in X$ and $\lambda > 0$ there exists some $\lambda' > 0$ such that $T(\lambda x) = \lambda' Tx$;

ascending (for p; cf. [11]) if there exists a continuous mapping φ of the unit interval $[0, 1]$ into itself with $\lambda < \varphi(\lambda)$ for $0 < \lambda < 1$ and such that for any $\lambda \in [0, 1]$ and any $x, y \in X$, $\lambda x \leqq y$ implies $\varphi(\lambda) Tx \leqq Ty$;

pointwise bounded (for p) if for every $x \in X$ there exist $u(x)$, $v(x) \in E_+$, $u(x) > 0$, such that $u(x) \leqq Tx \leqq v(x)$.

A sequence $(T_n)_n$ of operators $T_n: E_+ \to E_+$ is *uniformly ascending*, if all operators are ascending with the same φ, i.e., $\lambda x \leqq y$ implies $\varphi(\lambda) T_n x \leqq T_n y$ for all n. Similarly a sequence is *uniformly pointwise bounded* if $u(x) \leqq T_n x \leqq v(x)$ for all n.

On the unit level set X of the scale we now consider Hilbert's projective pseudometric, or *Hilbert's metric* for short (cf. [1], [2], [4], [8], [10], [11], [12], [14]). For $x, y \in E_+ \backslash \{0\}$ let $\lambda(x, y) = \sup\{\lambda \in \mathbb{R}_+ \mid \lambda x \leqq y\}$ and let $\mu(x, y) = -\log[\lambda(x, y) \cdot \lambda(y, x)]$. It is easily verified that μ is a metric on X except that μ may take on the value $+\infty$. The following lemma translates properties of positive operators into properties of operators acting on the metric space (X, μ).

LEMMA 1. *For an operator $T: E_+ \to E_+$ which is proper let $\tilde{T}: X \to X$, $\tilde{T}x = Tx/p(Tx)$ for $x \in X$.*

(i) *T subhomogeneous $\Rightarrow$ $\tilde{T}$ nonexpansive (on (X, μ)).*

(ii) *S proper and ray-preserving $\Rightarrow \widetilde{S \cdot T} = \tilde{S} \cdot \tilde{T}$.*

(iii) *T ascending $\Rightarrow \mu(\tilde{T}x, \tilde{T}y) \leqq c(x, y)$ for all $x, y \in X$. Thereby $c(x, y) = -\log[\varphi(\lambda(x, y)) \cdot \varphi(\lambda(y, x))]$ (φ as in the definition of "ascending") is continuous on $\{(x, y) \in X \times X \mid x > 0, y > 0\}$ with respect to Euclidean topology. Furthermore, $c(x, y) \leqq \mu(x, y)$ for all $x, y \in X$ and $c(x, y) < \mu(x, y)$ if $x \neq y$ and $\mu(x, y) < +\infty$.*

(iv) *T pointwise bounded and subhomogeneous $\Rightarrow$ There exists $a, b \in X$, $a \leqq \tilde{T}x \leqq b$ for all $x \in X$. If $(T_n)_n$ is uniformly pointwise bounded, then the bounds a, b are independent of n.*

Proof. (i) Since T is subhomogeneous it follows that $\lambda(\tilde{T}x, \tilde{T}y) \geqq (p(Tx)/p(Ty))\lambda(x, y)$ for all $x, y \in X$. Hence by the definition of Hilbert's metric $\mu(\tilde{T}x, \tilde{T}y) \leqq \mu(x, y)$, i.e., $\tilde{T}$ is nonexpansive on (X, μ).

(ii) With S, T proper, $S \cdot T$ is proper too. Since S is ray-preserving, for a given $x \in X$ there exists $\lambda' > 0$ such that $S(Tx/p(Tx)) = \lambda' S(Tx)$. Hence

$$\tilde{S}(\tilde{T}x) = \frac{S(\tilde{T}x)}{p(S(\tilde{T}x))} = \frac{\lambda' S \cdot T(x)}{p(\lambda' S \cdot T(x))} = \frac{S \cdot T(x)}{p(S \cdot T(x))} = \widetilde{S \cdot T}(x).$$

(iii) Applying p to $\lambda x \leqq y$ it follows that $\lambda(x, y) \in [0, 1]$ for $x, y \in X$. Since T is ascending it follows that $\varphi(\lambda(x, y))Tx \leqq Ty$, and hence $\lambda(\tilde{T}x, \tilde{T}y) \geqq (p(Tx)/p(Ty))\varphi(\lambda(x, y))$.

Interchanging the roles of x and y the definition of μ yields $\mu(\tilde{T}x, \tilde{T}y) \leqq c(x, y)$ with $c(x, y)$ as stated in the assertion. Suppose $x, y \in X$, $x \neq y$, and $\mu(x, y) < +\infty$. Because of the latter, $0 < \lambda(x, y)$ and $0 < \lambda(y, x)$. $\lambda(x, y) \leqq 1$ and $\lambda(y, x) \leqq 1$, because of $x, y \in X$, and equality in both cases would imply $x \leqq y$ and $y \leqq x$, i.e., $x = y$. Without restriction we may assume $\lambda(x, y) < 1$. Therefore by the properties of φ $\lambda(x, y) \cdot \lambda(y, x) < \varphi(\lambda(x, y)) \cdot \varphi(\lambda(y, x))$, implying $c(x, y) < \mu(x, y)$. From this $c(x, y) \leqq \mu(x, y)$ for $x, y \in X$. Finally, $c(x, y)$ depends continuously on $x > 0$, $y > 0$ since by an easy calculation $\lambda(x, y) = \min\{y_i/x_i \mid i \in \{1, \cdots, k\}\}$.

(iv) We shall show $u \leqq Tx \leqq v$ with $u, v \in E_+$, $u > 0$. Then $p(u) \leqq p(Tx) \leqq p(v)$, and (iv) follows by setting $a = u/p(v)$, $b = v/p(u)$. Let T be subhomogeneous and $u(x) \leqq Tx \leqq v(x)$ for $x \in X$, $u(x), v(x) \in E_+$, $u(x) > 0$. Denote by e_i, $i = 1, \cdots, k$, the vector in $E = \mathbb{R}^k$ having 1 in component i and 0's otherwise and let $e = e_1 + \cdots + e_k$. Define $u = (\min_i p(e_i)/p(e)) \min_i u(e_i/p(e_i))$, where $\min_i u(e_i/p(e_i))$ is a vector the jth component of which is obtained by taking the minimum over i of the jth component of $u(e_i/p(e_i))$. Define $v = (p(e)/\min_i p(e_i))v(e/p(e))$. Obviously, $u, v \in E_+$ and $u > 0$. To see $u \leqq Tx \leqq v$, let $x \in X$. If x_i denotes the ith component of x, then $x_i e_i \leqq x \leqq (\max_i x_i)e$ and by applying p $x_i p(e_i) \leqq p(x) = 1 \leqq (\max_i x_i)p(e)$ and hence $1/p(e) \leqq \max_i x_i \leqq 1/\min_i p(e_i)$. Since T is subhomogeneous, from $(\min_i p(e_i)/p(e))x \leqq e/p(e)$ it follows that $(\min_i p(e_i)/p(e))Tx \leqq T(e/p(e)) \leqq v(e/p(e))$. According to the definition of v, therefore, $Tx \leqq v$. Furthermore, from $x_i p(e_i)(e_i/p(e_i)) \leqq x$ subhomogeneity of T yields $x_i p(e_i) T(e_i/p(e_i)) \leqq Tx$. From the definition of u it follows that

$$p(e)x_i u \leqq x_i p(e_i) u\left(\frac{e_i}{p(e_i)}\right) \leqq x_i p(e_i) T\left(\frac{e_i}{p(e_i)}\right) \leqq Tx.$$

Hence $(\max_i x_i)p(e)u \leqq Tx$ and $u \leqq Tx$ because of $(\max_i x_i)p(e) \geqq 1$. Finally, the independence statement holds by construction of a and b. □

To apply the results of the previous section, we need the following comparison of Hilbert's metric and the maximum metric $|x - y| = \max_i |x_i - y_i|$ as defined on $\mathbb{R}^k$. (x_i the ith component of x.)

LEMMA 2. *For any $x, y \in X$*

$$(\min_i x_i)[1 - \exp(-\tfrac{1}{2}\mu(x, y))] \leqq |x - y| \leqq \max_i \{x_i, y_i\}(1 - \exp(-\mu(x, y))).$$

Proof. To see the first inequality, let $r = \min_i x_i$. If $r \leqq |x - y|$, then the first inequality holds trivially. Suppose $r > |x - y|$. Obviously $r(x - y) \leqq |x - y|x$, and hence $(1 - (|x - y|/r))x \leqq y$. Thus $\lambda(x, y) \geqq 1 - (|x - y|/r)$ and $\lambda(x, y) \cdot \lambda(y, x) \geqq (1 - (|x - y|/r))^2$. Because of $\lambda(x, y) \cdot \lambda(y, x) = \exp(-\mu(x, y))$ this proves the first inequality. For the second inequality observe that $x - y \leqq (1 - \lambda(x, y))x \leqq (1 - \lambda(x, y)\lambda(y, x))x$ and therefore $x_i - y_i \leqq (1 - \lambda(x, y)\lambda(y, x))x_i$. By interchanging the roles of x and y and taking the maximum over i the second inequality is obtained. □

Remark. The proof shows that the first inequality is true for any $x, y \in E_+ \setminus \{0\}$.

Our first application is concerned with weak ergodicity, a concept which originally stems from the theory of Markov chains (cf. [4], [8], [13], [14]). More generally, there

holds *weak ergodicity* (relative to some fixed scale p) for a sequence $(T_n)_n$ of operators on $\mathbb{R}_+^k$, $T_n:\mathbb{R}_+^k\to\mathbb{R}_+^k$, whenever for arbitrary nonzero starting points $x_1, y_1\in\mathbb{R}_+^k$ and x_n, y_n defined by $x_n = T_{n-1}\cdot\cdots\cdot T_2\cdot T_1(x_1)/p(T_{n-1}\cdot\cdots\cdot T_2T_1(x_1))$, y_n analogously with y_1 instead of x_1, the sequence $x_n - y_n$ tends to 0 for $n\to\infty$ in the Euclidean topology.

THEOREM 4 (*Weak ergodicity for nonlinear operators*). *Let $(T_n)_n$ be a sequence of operators on $\mathbb{R}_+^k$ that are proper, subhomogeneous, and ray-preserving. Suppose there is some $r\geqq 1$ such that the sequence of lumped operators $(S_m)_m$ defined by $S_m = T_{m+r-1}\cdot\cdots\cdot T_{m+1}\cdot T_m$ is uniformly ascending and uniformly pointwise bounded. Then there holds weak ergodicity for the sequence $(T_n)_n$.*

Proof. The theorem will be a consequence of Theorem 1. Let $X=\{x\in\mathbb{R}_+^k \mid p(x)=1\}$ and put $f_n=\tilde{T}_n$. By Lemma 1(ii) $\tilde{S}_m=\tilde{T}_{m+r-1}\cdot\cdots\cdot\tilde{T}_{m+1}\cdot\tilde{T}_m=f_{m+r-1}\cdot\cdots\cdot f_{m+1}\cdot f_m=F_m$, F_m being a lumped operator for $(f_n)_n$. Since $(S_m)_m$ is assumed to be uniformly ascending, each S_m has to be subhomogeneous. Hence by the assumption of uniform pointwise boundedness from Lemma 1(iv) the existence of $a, b\in\mathbb{R}_+^k$ $a>0$ follows such that $F_m(X)\subseteq Y$ for $Y=\{x\in X \mid a\leqq x\leqq b\}$. Because of $a>0$, the first inequality of Lemma 2 yields $s=\sup\{\mu(x,y)\mid x,y\in Y\}<+\infty$. Truncating μ as $d(x,y)=\min(\mu(x,y), s)$ for $x, y\in X$ we obtain the metric space (X, d). Since Y is compact in the Euclidean topology (p was assumed to be continuous) according to Lemma 2, Y is compact also in (X, d). f_n is nonexpansive on (X, d) because of Lemma 1(i). Furthermore, by Lemma 1(iii) $\mu(F_m(x), F_m(y))\leqq c(x,y)$ for all $x, y\in X$ and $c(x,y)<d(x,y)$ for all $x\neq y$ with $\mu(x,y)<+\infty$ where $c(x,y)=-\log[\varphi(\lambda(x,y))\cdot\varphi(\lambda(y,x))]$. Thus we obtain for all $x, y\in Y$ $d(F_m(x), F_m(y))\leqq c(x,y)$ and $c(x,y)<d(x,y)$ provided $x\neq y$. That is, $(F_m)_m$ is contractive on Y and we may apply Theorem 1. By this theorem together with Lemma 2 $x_n-y_n\to 0$ in the Euclidean topology, whereby

$$x_n = f_{n-1}\cdot\cdots\cdot f_2\cdot f_1(x_1)=\frac{T_{n-1}\cdot\cdots\cdot T_2\cdot T_1(x_1)}{p(T_{n-1}\cdot\cdots\cdot T_2\cdot T_1(x_1))},$$

and y_n analogously. The starting points x_1, y_1 are arbitrary in X, but since the T_i are ray-preserving we may allow for arbitrary nonzero starting points in all of $\mathbb{R}_+^k$. □

An interesting special case of the theorem is obtained if concave operators are considered. An operator $T:\mathbb{R}_+^k\to\mathbb{R}_+^k$ is *concave* whenever $T(\lambda x+(1-\lambda)y)\geqq\lambda Tx+(1-\lambda)Ty$ for any $x, y\in\mathbb{R}_+^k$ and any $\lambda\in[0,1]$.

COROLLARY (*Concave version of the Coale-Lopez theorem*). *Consider a scale induced on $\mathbb{R}_+^k$ by a vector space norm. There holds weak ergodicity for every sequence $(T_n)_n$ of proper, ray-preserving and concave operators on $\mathbb{R}_+^k$ provided some sequence of lumped operators $S_m = T_{m+r-1}\cdot\cdots\cdot T_{m+1}\cdot T_m$ is uniformly pointwise bounded.*

Proof. To obtain the corollary from Theorem 4 we show that T_n is subhomogeneous and that $(S_m)_m$ is ascending. Consider a concave operator T on $\mathbb{R}_+^k$ and let $\lambda x\leqq y$ for $x, y\in X$ and $0\leqq\lambda<1$. For $z=y-\lambda x$, $z\geqq 0$ and $y=\lambda x+(1-\lambda)(z/(1-\lambda))$. Concavity implies $Ty\geqq\lambda Tx+(1-\lambda)T(z/(1-\lambda))$. In particular $Ty\geqq\lambda Tx$ which by approximation is seen to be true also for $\lambda=1$. Hence each T_n is subhomogeneous. Suppose now for all $x\in X$ and some $u, v\in\mathbb{R}_+^k$, $u>0$, $u\leqq Tx\leqq v$. By subhomogeneity of T

$$T\left(\frac{z}{1-\lambda}\right)=T\left(\frac{p(z)}{1-\lambda}\,\frac{z}{p(z)}\right)\geqq\frac{p(z)}{1-\lambda}\,T\left(\frac{z}{p(z)}\right)$$

and therefore $Ty\geqq\lambda Tx+p(z)T(z/p(z))$. Since p is induced by a norm, $p(y)=p(\lambda x+z)\leqq\lambda p(x)+p(z)$ and hence $p(z)\geqq 1-\lambda$. Furthermore, there is a real number $0<s\leqq 1$ such that $sv\leqq u$ and therefore $sTx\leqq sv\leqq u\leqq T(z/p(z))$. Thus we obtain $Ty\geqq\lambda Tx+(1-\lambda)sTx=\varphi(\lambda)Tx$, with $\varphi(\lambda)=\lambda+(1-\lambda)s$. This formula is true by approximation also for $\lambda=1$. Being a composite of concave operators S_m is concave

too and by the uniform pointwise b undedness there exist $u, v \in \mathbb{R}_+^k$, $u > 0$, such that $u \leqq S_m x \leqq v$ for all $x \in X$ and all m. Putting $T = S_m$ we conclude that $\lambda x \leqq y$ for $x, y \in X$ and $\lambda \in [0, 1]$ implies $\varphi(\lambda) S_m x \leqq S_m(y)$ for all m, whereby $\varphi(\lambda) = \lambda + (1-\lambda)s$. This shows that $(S_m)_m$ is ascending. □

The weak ergodicity theorem or Coale-Lopez theorem referred to in the literature (cf. [4], [8], [13], [14]) is contained in the above corollary as the special case of linear operators T_n. In that special case weak ergodicity holds provided the T_n are proper and $(S_m)_m$ is uniformly pointwise bounded. For example, this is guaranteed if all S_m are strictly positive (for some $r \geqq 1$) and all the possible entries (all the possible nonzero entries) of the matrices T_n for $n = 1, 2, \cdots$, are bounded from above (bounded from below by some positive constant) (cf. [8], [14]; in [8] weak ergodicity is not with respect to the Euclidean topology but with respect to the topology belonging to Hilbert's metric).

The concept of strong ergodicity also stems from the theory of Markov chains. Generalizing, we say there holds *strong ergodicity* (relative to some fixed scale p) for a sequence $(T_n)_n$ of operators on $\mathbb{R}_+^k$, when for arbitrary nonzero starting points $x_1 \in \mathbb{R}_+^k$ the sequence defined by

$$x_n = \frac{T_{n-1} \cdot \cdots \cdot T_2 \cdot T_1(x_1)}{p(T_{n-1} \cdot \cdots \cdot T_2 \cdot T_1(x_1))}$$

converges in the Euclidean topology to the same limit point x^*.

THEOREM 5 (*Strong ergodicity for nonlinear operators*). *Let $(T_n)_n$ and $(S_m)_m$ be sequences of operators on $\mathbb{R}_+^k$ as in Theorem 4. Suppose in addition for the lumped operators S_m uniform convergence on $X = \{x \in \mathbb{R}_+^k \mid p(x) = 1\}$ equipped with the Euclidean metric to some operator S on $\mathbb{R}_+^k$. Then there holds strong ergodicity for the sequence $(T_n)_n$ and the limit point x^* is the unique eigenvector of S in X.*

Proof. Putting $f_n = \tilde{T}_n$, $F_m = \tilde{S}_m$ the assumptions of Theorem 1 are satisfied according to the proof of Theorem 4. To apply Theorem 2 we show uniform convergence of F_m to $F = \tilde{S}$ on the metric space (X, d) with $d(x, y) = \min(\mu(x, y), s)$ being the truncated Hilbert metric (as in the proof of Theorem 4). Since by assumption $(S_m)_m$ is uniformly pointwise bounded it follows that $S_m x \geqq u > 0$ for all m, all $x \in X$ (as in the proof of part (iv) of Lemma 1). In particular $Sx \geqq u > 0$ for all $x \in X$ and $F = \tilde{S}$ is well defined on X. Because of Lemma 2 (and the remark thereafter) and because of $\mu(S_m x, Sx) = \mu(F_m(x), F(x))$ for $x \in X$ the uniform convergence of S_m to S for the Euclidean metric implies uniform convergence of F_m to F for the metric d. Theorem 2 yields convergence with respect to d of

$$x_n = f_{n-1} \cdot \cdots \cdot f_2 \cdot f_1(x_1) = \frac{T_{n-1} \cdot \cdots \cdot T_2 \cdot T_1(x_1)}{p(T_{n-1} \cdot \cdots \cdot T_2 \cdot T_1(x_1))}, \qquad x_1 \in X,$$

to the unique fixed point x^* of F in X. By Lemma 2 convergence is also with respect to Euclidean topology. Obviously fixed points of F in X correspond in a unique manner to eigenvectors of S in X. Because the T_n are ray-preserving, $x_1 \in X$ may be replaced by any nonzero $x_1 \in \mathbb{R}_+^k$. This proves strong ergodicity of $(T_n)_n$. □

As for weak ergodicity we may specialize to concave operators.

COROLLARY (*Strong ergodicity for concave operators*). *Consider a scale induced on $\mathbb{R}_+^k$ by a vector space norm. There holds strong ergodicity for every sequence $(T_n)_n$ of proper, ray-preserving, and concave operators on $\mathbb{R}_+^k$ provided some sequence of lumped operators $(S_m)_m$ is uniformly pointwise bounded and converges uniformly on $X = \{x \in \mathbb{R}_+^k \mid p(x) = 1\}$ equipped with the Euclidean metric to some operator S on $\mathbb{R}_+^k$.*

Proof. $(T_n)_n$, $(S_m)_m$ satisfy the assumptions of Theorem 4, according to the proof of the corollary following Theorem 4. Hence the above corollary is implied by Theorem 5. □

Using Theorem 3 of the previous section, we may obtain sufficient conditions for strong ergodicity without referring to lumped operators. For this from matrix theory we borrow the following notion. An operator $T:\mathbb{R}_+^k \to \mathbb{R}_+^k$ is *primitive* (for p), if there exists $r \geqq 1$ such that for any

$$x, y \in X = \{x \in \mathbb{R}_+^k \mid p(x) = 1\} \quad \text{and} \quad \lambda \in \mathbb{R}_+ \quad \lambda x \leqq y, \lambda x \neq y \quad \text{implies that } \lambda T^r x < T^r y.$$

THEOREM 6. *Let $T_n:\mathbb{R}_+^k \to \mathbb{R}_+^k$, $n = 1, 2, \cdots$, be proper, subhomogeneous, and ray-preserving operators that converge, uniformly on the intersection with the unit sphere of some monotonic norm $\|\cdot\|$, to an operator T on $\mathbb{R}_+^k$. Suppose the operator T is proper ray-preserving, continuous, and primitive (for $\|\cdot\|$). Then there holds strong ergodicity for $(T_n)_n$ with scale $\|\cdot\|$ and the limit point x^* is the unique normalized eigenvector of T.*

Proof. $X = \{x \in \mathbb{R}_+^k \mid \|x\| = 1\}$ equipped with $e(x, y) = \|x - y\|$ is a compact metric space on which $(T_n)_n$ converges uniformly to an operator T which must be uniformly continuous. Formula $(*)$ of step (3) in the proof of Theorem 2 then yields that on (X, e) the sequence of lumped operators $S_m = T_{m+r-1} \cdot \cdots \cdot T_{m+1} \cdot T_m$ converges uniformly to T^r (r as in the definition of primitivity). Take $\|\cdot\|$ as scale and put $f_n = \tilde{T}_n$, $f = \tilde{T}$. From Lemma 1 for the lumped operators F_m to $(f_n)_n$ $F_m = \tilde{S}_m$ and $F = \tilde{T}^r = \tilde{T}^r = f^r$. Since T is primitive, T^r is in particular pointwise bounded and subhomogeneous (using continuity of T). Hence by Lemma 1(iv) there exist $a, b \in X = \{x \in \mathbb{R}_+^k \mid \|x\| = 1\}$, $a > 0$ such that $a \leqq f^r(x) \leqq b$. From Lemma 2 (and the remark thereafter) it follows that on X F_m converges uniformly to F with respect to Hilbert's metric μ. We shall apply Theorem 3 for $U = \{y \in X \mid \frac{1}{2}a < y < 2b\}$. Obviously, U is relatively open in X for the Euclidean topology and the closure $\bar{U} = \{y \in X \mid \frac{1}{2}a \leqq y \leqq 2b\}$ is compact. For $d(x, y) = \min(\mu(x, y), s)$, $s = \sup\{\mu(x, y) \mid x, y \in \bar{U}\} < +\infty$ from Lemma 2 it follows that U is open in (X, d), the closure of U in this space equals $\bar{U}$ and $\bar{U}$ is compact in (X, d). Furthermore $F(X) \subseteq U$. Finally, from the primitivity of T it follows (as for part (iii) of Lemma 1) that $\mu(F(x), F(y)) < \mu(x, y)$ for $x, y \in \bar{U}$. Hence F is contractive on $\bar{U}$ for d.

Thus Theorem 3 yields the convergence relative to d of the sequence defined by $x_{n+1} = f_n(x_n)$, $x_1 \in X$, to some $x^* \in X$. Also, x^* is the unique fixed point of F in X. Primitivity of T implies $x^* = T^r x^* / \|T^r x^*\| > 0$. By Lemma 2 $(x_n)_n$ converges to x^* in the Euclidean topology. Since the T_n are ray-preserving, any nonzero $x_1 \in \mathbb{R}_+^k$ is allowed to be a starting point. Thus $(T_n)_n$ is strongly ergodic. Concerning the assertion on x^* in the theorem, it suffices to show x^* to be an eigenvector of T. Since $x^* > 0$, there exists $0 < \lambda$ sufficiently small with $\lambda T^{r-1} x^* / \|T^{r-1} x^*\| \leqq x^*$. Being the limit of subhomogeneous operators, T is subhomogeneous and therefore $\lambda T(T^{r-1} x^* / \|T^{r-1} x^*\|) \leqq Tx^*$. Since T is ray-preserving it follows that $\mu T^r x^* \leqq Tx^*$ for some $\mu > 0$ and hence $Tx^* > 0$. Because of $T_n x_n \to Tx^*$ in the Euclidean topology by Lemma 2 it follows that $f_n(x_n) \to f(x^*)$ for d. But $f_n(x_n) = x_{n+1} \to x^*$, and hence $f(x^*) = x^*$, i.e., x^* is an eigenvector of T. □

As a simple special case the theorem on strong ergodicity for nonnegative matrices (cf. [14]) follows directly from Theorem 6. For T_n linear the only assumptions placed on T_n by Theorem 6 are properness (e.g., fulfilled if there is no column of zeros) and pointwise convergence of T_n to an operator T some power of which is strictly increasing.

4. Example: A density- and time-dependent Leslie model. Consider a population of female individuals at discrete points of time. The growth of the population is

dependent upon birth and death which are, among others, age-specific. Hence the population is divided into a finite number of age groups, say $1, 2, \cdots, k$. Denote by $X_i(t)$ *the number of individuals in age group* $i \in \{1, \cdots, k\}$ *at time* $t \in \{1, 2, \cdots\}$ and let $X(t) = (X_1(t), \cdots, X_k(t))$ be the *population vector at time t.* The *total population at time t* is given by $\|X(t)\|$, where $\|\cdot\|$ is the norm on $\mathbb{R}^k$ defined by $\|x\| = \Sigma|x_i|$. The vector $x(t) = X(t)/\|X(t)\|$ is called the *age structure at time t.* The *birth rate in age group i* is denoted by $b_i(t, X(t))$ and may depend on time and the population vector. The *survival rate in age group i,* denoted by $s_i(t, X(t))$, specifies the proportion of $X_i(t)$ surviving to be in age group $i+1$ at time $t+1$. The survival rate $s_k(\cdot, \cdot)$ for the oldest group therefore is 0. It is assumed that all the other survival rates and the birth rates are strictly positive. Thus we arrive at the following *model*:

$$X_1(t+1) = \sum_{i=1}^{k} b_i(t, X(t))X_i(t),$$

$$X_{i+1}(t+1) = s_i(t, X(t))X_i(t) \quad \text{for } i = 1, \cdots, k-1.$$

Or, equivalently,

$$X(t+1) = T(t)X(t) \quad \text{where } T(t): \mathbb{R}_+^k \to \mathbb{R}_+^k$$

is given by

$$T(t)x = L(t, x)x = \begin{bmatrix} b_1(t, x) & \cdots & b_k(t, x) \\ s_1(t, x) & 0 & 0 \\ 0 & & \vdots \\ \vdots & \ddots & \\ 0 & s_{k-1}(t, x) & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}.$$

This is a *generalized Leslie model* where birth rates and survival rates are allowed to depend on time and on density $X(t)$. In the original *Leslie model* (cf. [4], [8], [13], [14]) birth rates and survival rates are assumed to be constant, and hence the *Leslie matrix* $L(t, x)$ is constant, $L(t, x) = L$ for all t and x. Since L turns out to be a primitive matrix this case is already covered by a theorem on matrices due to Perron which implies that the age structure $x(t)$ converges to an equilibrium x^* (cf. [12], [14]).

The case that $L(t, x)$ is time-dependent only, $L(t, x) = L(t)$, is covered by the (linear) Coale-Lopez theorem which yields weak ergodicity of the age structure $x(t)$ (cf. [4], [8], [13], [14]). However, if $L(t, x)$ also depends on x and therefore $T(t)$ becomes a nonlinear operator, a lot of things may happen. So it is shown in [9] for the very simple model given by $k = 2$, $b_i(t, x) = b_i \exp[-a(x_1 + x_2)]$, $s_1(t, x) = \text{constant}$, that chaotic dynamics occurs for certain choices of the parameters b_1, b_2, a.

Being interested in weak ergodicity also for the nonlinear case we therefore have to make some assumptions which constitute our concave Leslie model.

The concave Leslie model is based on the following assumptions.

(1) The functions $x \to b_i(t, x)x_i$, $x \to s_i(t, x)x_i$ are concave on $\mathbb{R}_+^k$ for all i, t.

(2) $b_i(t, \lambda x)/b_i(t, x) = s_j(t, \lambda x)/s_j(t, x)$ for all $\lambda > 0$ and all i, j, x.

(3) There exist functions $c(\cdot)$, $d(\cdot): \mathbb{R}_+^k \backslash \{0\} \to \mathbb{R}_+ \backslash \{0\}$ with $c(\cdot)$ not increasing, $d(\cdot)$ not decreasing with respect to "$\leqq$" such that

$$c(x) \leqq b_i(t, x), s_i(t, x) \quad \text{and} \quad b_i(t, x)x_i, s_i(t, x)x_i \leqq d(x)$$

for all i, t and x.

We show that the concave Leslie model satisfies the assumptions of the concave version of the Coale-Lopez theorem presented in the last section. Assumption (1)

obviously implies that all the operators $T(t)=L(t,x)x$ are concave. The assumption itself means that the number of births or survivals contributed by a particular age group decreases by "population pressure" with the density in that group. Assumption (2) implies $L(t,\lambda x)=\mu L(t,x)$ for some μ, which may depend on λ, t, x and hence every $T(t)$ is ray-preserving. By this assumption a certain "homogeneity" of the vitality rates with respect to changes in total population is required. The assumption is met, e.g., if all vitality rates are homogeneous of the same degree with respect to total population. The assumption (3) requiring uniform upper and lower bounds for the vitality rates implies obviously that the $T(t)$ are proper and yields the existence of a uniformly pointwise bounded sequence of inhomogeneous iterates as shown by the following lemma.

LEMMA 3. *By assumption* (3) *there exist for every* $x\in\mathbb{R}^k_+\backslash\{0\}$ $u(x)$, $v(x)\in\mathbb{R}^k_+$ *where* $u(x)>0$ *such that*

$$u(x)\leqq T(m+k-1)\cdot\cdots\cdot T(m+1)\cdot T(m)x\leqq v(x)\quad \textit{for all } m\in\mathbb{N}.$$

Proof. By (3) $T(t)x\leqq w(x)$ for all t, where $w(x)=(kd(x),d(x),\cdots,d(x))$ and $w(\cdot)$ is not decreasing with respect to "$\leqq$." Hence $T(t)T(s)x\leqq w(T(s)x)\leqq w\cdot w(x)$. By iteration, if $S(m)=T(m+k-1)\cdot\cdots\cdot T(m+1)\cdot T(m)$, $S(m)x\leqq w\cdot\cdots\cdot w(x)$. Choose $v(x)=w^k(x)=w\cdot\cdots\cdot w(x)$. Furthermore by (3) $L(t,x)\geqq c(x)L$, where

$$L=\begin{bmatrix}1&1&\cdots&1\\1&0&\cdots&0\\0&&&\vdots\\\vdots&&\ddots&\vdots\\0&0&1&0\end{bmatrix}.$$

Hence $T(t)x=L(t,x)x\geqq c(x)Lx$ for all t. It follows, by using $T(s)x\leqq w(x)$, that

$$T(t)T(s)x\geqq c(T(s)x)LT(s)x\geqq c(T(s)x)c(x)L^2x\geqq c(w(x))c(x)L^2x.$$

Therefore by iteration for $S(k)$ defined above

$$S(m)x\geqq c(w^{k-1}(x))\cdot c(w^{k-2}(x))\cdots c(w(x))c(x)L^kx.$$

Choose as $u(x)$ the right-hand side of this inequality. It is easily checked that the matrix L^k is strictly positive. Hence $u(x)>0$. □

Due to the lemma we can apply the concave version of the Coale–Lopez theorem (corollary to Theorem 4) which yields that weak ergodicity holds in the above concave Leslie model. A simple example of a concave Leslie model is the following one that contains the linear time-dependent Leslie model as a special case. Let for k arbitrary and $i\in\{1,\cdots,k\}$, $t\in\{1,2,\cdots\}$,

$$b_i(t,x)=b_i(t)x_i^{\alpha-1},\qquad s_i(t,x)=s_i(t)x_i^{\alpha-1},$$

where $0<\alpha\leqq 1$ and $c_1\leqq b_i(t)$, $s_i(t)\leqq c_2$ for all i, all t with certain positive constants c_i. Assumptions (1) and (2) for a concave Leslie model are obviously satisfied. Assumption (3) is fulfilled by choosing the following functions:

$$c(x)=c_1\min_i x_i^{\alpha-1},\qquad d(x)=c_2\sum_i x_i^{\alpha}\quad\text{for } x\in\mathbb{R}^k_+\backslash\{0\}.$$

The linear Leslie model is obviously contained for $\alpha=1$. For $\alpha<1$, e.g., $\alpha=\frac{1}{2}$, the operator $T(t)$ is neither positive homogeneous nor additive. Nevertheless, as for the linear case, weak ergodicity holds for this simple nonlinear example.

The example may serve also to illustrate the general Theorem 6 on strong ergodicity. For this suppose that for all i, $b_i(t)$ and $s_i(t)$ tend to $b_i > 0$ and $s_i > 0$, respectively, as $t \to \infty$. If

$$L(t) = \begin{bmatrix} b_1(t) & \cdots & b_k(t) \\ s_1(t) & & 0 \\ 0 & & \vdots \\ \vdots & & \\ 0 \quad \cdots & s_{k-1}(t) & 0 \end{bmatrix}, \qquad L = \begin{bmatrix} b_1 & \cdots & b_k \\ s_1 & & 0 \\ 0 & & \vdots \\ \vdots & & \\ 0 \quad \cdots & s_{k-1} & 0 \end{bmatrix},$$

$$x^\alpha = \begin{bmatrix} x_1^\alpha \\ x_2^\alpha \\ \vdots \\ x_k^\alpha \end{bmatrix},$$

then $T(t)x = L(t)x^\alpha$ and $T(t)$ converges uniformly on $X = \{x \in \mathbb{R}_+^k \mid \|x\| = 1\}$ to the operator T given by $Tx = Lx^\alpha$. Obviously, $T(t)$ and T are proper and T is continuous on X. $T(t)$ is also ray-preserving and subhomogeneous because of

$$T(t)(\lambda x) = L(t)(\lambda x)^\alpha = \lambda^\alpha L(t)x^\alpha = \lambda^\alpha T(t)x.$$

The same applies to T. To see primitivity let $\lambda x \leqq y$, $\lambda x \neq y$ for $x, y \in X$. Successive application of L yields, using the monotonicity of $x \mapsto x^\alpha$, that $T^k(\lambda x) < T^k y$. Hence $\lambda T^k x \leqq \lambda^{\alpha^k} T^k x = T^k(\lambda x) < T^k y$, and T is primitive. Thus Theorem 6 supplies strong ergodicity for this little nonlinear example, a result that still contains the corresponding result for the linear case.

REFERENCES

[1] G. BIRKHOFF, *Extensions of Jentzsch's theorem*, Trans. Amer. Math. Soc., 85 (1957), pp. 219-227.

[2] ———, *Lattice Theory*, 3rd ed., Coll. Publ. Ser. Vol. 25, American Mathematical Society, Providence, RI, 1967.

[3] F. E. BROWDER, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, Proc. Sympos. Pure Math., Vol. 18, Part 2, 1976.

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

[5] K. DEIMLING, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[6] T. FUJIMOTO AND U. KRAUSE, *Ergodicity for inhomogeneous products of nonlinear positive operators*, mimeo (1985).

[7] ———, *Ergodic price setting with technical progress*, in Competition, Instability, and Nonlinear Cycles, W. Semmler, ed., Springer-Verlag, Berlin, 1986.

[8] M. GOLUBITSKY, E. B. KEELER, AND M. ROTHSCHILD, *Convergence of the age structure: applications of the projective metric*, Theoret. Population Biol., 7 (1975), pp. 84-93; addendum ibid., 10 (1976), p. 413.

[9] J. GUCKENHEIMER, G. OSTER, AND A. IPAKTSCHI, *Dynamics of density dependent population models*, J. Math. Biol., 4 (1977), pp. 101-147.

[10] V. I. ISTRĂTESCU, *Fixed Point Theory*, D. Reidel, Dordrecht, 1981.

[11] U. KRAUSE, *A nonlinear extension of the Birkhoff-Jentzsch theorem*, J. Math. Anal. Appl., 114 (1986), pp. 552-568.

[12] ———, *Perron's stability theorem for non-linear mappings*, J. Math. Econom., 15 (1986), pp. 275-282.

[13] J. H. POLLARD, *Mathematical Models for the Growth of Human Populations*, Cambridge University Press, Cambridge, 1973.

[14] E. SENETA, *Non-Negative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, Berlin, 1980; 1st ed. Non-negative Matrices, G. Allen and Unwin, London, 1973.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

007

ANALYSIS OF LARGE DEFORMATION OF A HEAVY CANTILEVER*

SZE-BI HSU† AND SHIN-FENG HWANG‡

Abstract. In this paper a mathematical model is discussed describing the deformation of a cantilever by its own weight. We assume that a cantilever of uniform cross-section and density is held fixed at an angle α at one end and is free at the other end. The shape of the cantilever depends heavily on α and a nondimensional parameter K which represents the relative importance of density and length to that of flexural rigidity. We analyze the bifurcation phenomena for the vertical case, $\alpha = \pi$. Several numerical results are presented and discussed.

Key words. bifurcation, Sturm comparison, nonlinear eigenvalue problem, nonlinear oscillation

AMS(MOS) subject classifications. 73K05, 34B15, 34C10, 34C15

1. Introduction. The deformation of a cantilever by its own weight is of interest both practically due to its engineering significance and theoretically due to its inherent nonlinearity. We assume that a cantilever of uniform cross-section and density is held fixed at an angle α at one end and is free at the other end. If the cantilever is thin enough then its deformed shape can be described by the elastica theory. Using this approximation and small deflections, Euler first investigated the stability of a vertical cantilever (column) under its own weight [3]. Euler's stability problem was later corrected by Greenhill [4] who obtained the minimum unstable height for a column of given density and rigidity. The large deformation of a heavy elastica was first numerically integrated by Bickeley [1] who found only one of the solutions of the originally horizontal cantilever. Later, Wang [8] used the perturbation method on the elastica equations for a small and large parameter K, where K is a nondimensional parameter which represents the relative importance of density and length to that of flexural rigidity. In [8] Wang also studied the bifurcation phenomena numerically as the parameters K, α change.

In this paper we first give the uniqueness results for the solutions of the elastica equation. We then give the complete bifurcation results for the vertical case, $\alpha = \pi$, in the spirit of [6], [7]. From these analytic results we improve the numerical results in [8] and give the reliable numerical computation results.

2. Formulation. We assume a cantilever of uniform density ρ and total length L, is held fixed at an angle α at one end, say, the origin, and is free at the other end. Let us consider a small segment of the cantilever. A moment balance gives (see Fig. 1)

$$m - \rho(L - s') \sin\theta \, ds' = m + dm, \tag{2.1}$$

where $m = m(s')$ is the local moment, s' is the arc length from the origin, and $\theta = \theta(s')$ is the local angle of inclination. According to Euler, the local moment is proportional to the curvature $d\theta/ds'$, i.e.,

$$m = -\mathrm{EI}\frac{d\theta}{ds'}, \tag{2.2}$$

* Received by the editors August 6, 1986; accepted for publication (in revised form) June 9, 1987.

† Institute of Applied Mathematics, Tsing-Hua University, Hsinchu, Taiwan, Republic of China. The research of this author was supported in part by National Research Council, Republic of China.

‡ Department of Applied Mathematics, Chiao-Tung University, Hsinchu, Taiwan, Republic of China.

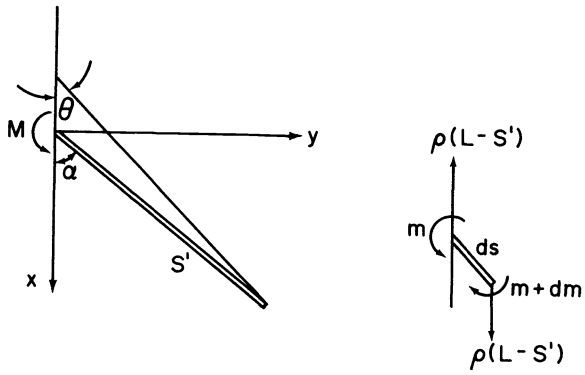


FIG. 1

where EI is the flexural rigidity of the material. From (2.1), (2.2), we obtain

$$\text{EI}\frac{d^2\theta}{ds'^2}=\rho(L-s')\sin\theta, \tag{2.3}$$

and the boundary conditions are

$$\theta(0)=\alpha, \qquad \frac{d\theta}{ds'}(L)=0. \tag{2.4}$$

Let $s=s'/L$ and then (2.3), (2.4) become

$$\begin{aligned}&\frac{d^2\theta}{ds^2}=K^3(1-s)\sin\theta, \quad K>0, \quad 0\leqq s\leqq 1,\\ &\theta(0)=\alpha, \quad \theta'(1)=0, \quad -\pi<\alpha<\pi.\end{aligned} \tag{2.5}$$

The important parameter $K=(\rho L^3/\text{EI})^{1/3}$ represents the relative importance of density and length to that of flexural rigidity.

The main concern of this paper is to determine the multiplicities of solutions of (2.5) provided that $K>0$, $-\pi\leqq\alpha\leqq\pi$ are given.

First of all, we shall reformulate our problem (2.5). Let

$$\psi(s)=\theta(1-s), \qquad 0\leqq s\leqq 1.$$

Then (2.5) becomes

$$\begin{aligned}&\frac{d^2\psi}{ds^2}=K^3 s\sin\psi, \quad 0<s<1, \quad K>0,\\ &\psi'(0)=0, \quad \psi(1)=\alpha, \quad -\pi\leqq\alpha\leqq\pi.\end{aligned} \tag{P}_\alpha$$

Since $\psi(s)$, $0\leqq s\leqq 1$, is a solution of $(\text{P})_\alpha$ if and only if $-\psi(s)$ is a solution of $(\text{P})_{-\alpha}$. Hence we only consider the problem with $0\leqq\alpha\leqq\pi$. We may also reduce the problem $(\text{P})_\alpha$, $0\leqq\alpha\leqq\pi$, by the following scaling:

$$\Psi(s)=\psi(s/K).$$

Then $\Psi(s)$ satisfies

$$\begin{aligned}&\frac{d^2\Psi}{ds^2}=s\sin\Psi(s), \qquad 0\leqq s\leqq K,\\ &\Psi'(0)=0, \quad \Psi(K)=\alpha, \quad 0<\alpha<\pi.\end{aligned} \tag{2.6}$$

3. Uniqueness of solutions of $(P)_\alpha$, $0 \leqq \alpha \leqq \pi$. In this section, we present some results concerning the uniqueness of solutions of boundary value problem

$$(P)_\alpha \qquad \frac{d^2\psi}{ds^2} = K^3 s \sin\psi, \quad 0 \leqq s \leqq 1, \quad K > 0,$$
$$\psi'(0) = 0, \quad \psi(1) = \alpha, \quad 0 \leqq \alpha \leqq \pi.$$

LEMMA 3.1. *The problem* $(P)_0$ *has a unique solution, namely,*

$$\psi(s) \equiv 0, \quad 0 \leqq s \leqq 1 \quad \textit{for any } K > 0.$$

Proof. Obviously $\psi(s) \equiv 0$ is a solution of $(P)_0$. Multiplying the equation in $(P)_0$ by $d\psi/ds$ and integrating the resulting equation from 0 to 1, we obtain

$$\frac{1}{2}(\psi'(1))^2 = K^3\left[\int_0^1 \cos\psi(s)\, ds - 1\right] \geqq 0.$$

However,

$$\int_0^1 \cos\psi(s)\, ds - 1 \leqq 0.$$

Hence we have $\psi'(1) = 0$. Since $\psi(1) = 0$, $\psi'(1) = 0$, the conclusion $\psi(s) \equiv 0$ follows directly from the uniqueness of solutions of ordinary differential equations. □

The existence of solution of problem $(P)_\alpha$ follows directly from the results in [2] since the right-hand side of $(P)_\alpha$, $K^3 s \sin\psi$, is a bounded function for $0 \leqq s \leqq 1$.

We now present a result concerning the uniqueness of solution of $(P)_\alpha$.

LEMMA 3.2. *If* $K^3 < \sqrt{45}$, *then* $(P)_\alpha$ *has a unique solution for every* $\alpha \in [0, \pi]$.

Proof. Let $\psi(s)$ be a solution of $(P)_\alpha$; then

$$\psi(s) = \alpha - \int_0^1 K^3 \xi \sin\psi(\xi) G(s, \xi)\, d\xi,$$

where

$$G(s, \xi) = 1 - \max(s, \xi).$$

Let $\psi_1(s)$, $\psi_2(s)$ be solutions of $(P)_\alpha$. Then

$$|\psi_1(s) - \psi_2(s)| \leqq \int_0^1 G(s, \xi)\xi |\psi_1(\xi) - \psi_2(\xi)|\, d\xi$$
$$\leqq K^3\left[\int_0^1 G^2(s, \xi)\xi^2\, d\xi\right]^{1/2} \|\psi_1 - \psi_2\|_2,$$

or

$$\|\psi_1 - \psi_2\|_2^2 = \int_0^1 |\psi_1(s) - \psi_2(s)|^2\, ds$$
$$\leqq K^6\left(\int_0^1 \int_0^1 G^2(s, \xi)\xi^2\, d\xi\right)\|\psi_1 - \psi_2\|_2^2$$

since

$$\int_0^1 \int_0^1 G^2(s, \xi)\xi^2\, d\xi\, ds = \frac{1}{45}.$$

If $K^6/45 < 1$ or $K^3 < \sqrt{45}$ then we must have

$$\psi_1 \equiv \psi_2. \qquad \square$$

4. The multiplicities of the solutions of $(P)_\alpha$ for $\alpha = \pi$. In this section we shall present the analytic results for the vertical case, $\alpha = \pi$. The analytic results for this special case will help us to understand the bifurcation phenomena for the general problem $(P)_\alpha$. In the rest of this section, we shall restrict our attention to the vertical case, $\alpha = \pi$:

$$\frac{d^2\psi}{ds^2} = K^3 s \sin\psi, \quad \psi'(0) = 0, \quad \psi(1) = \pi. \tag{4.1}$$

Let $s = x$, $v(x) = \psi(x/K) - \pi$. Then (4.1) takes the form

$$\begin{aligned} &v''(x) + x \sin v = 0, \qquad ' = d/dx, \\ &v'(0) = 0, \qquad v(K) = 0. \end{aligned} \tag{4.2}$$

We shall study the boundary value problem (4.2) by the shooting method and consider the following initial value problem

$$\begin{aligned} &v''(x) + x \sin v = 0, \\ &v'(0) = 0, \\ &v(0) = a, \qquad a \in \mathbb{R}. \end{aligned} \tag{4.3}$$

We denote the solution of (4.3) by $v(x, a)$. From the uniqueness of solutions of ordinary differential equations, it follows that

$$\begin{aligned} &v(x, 2\pi + a) = 2\pi + v(x, a), \\ &v(x, 2\pi - a) = 2\pi - v(x, a), \\ &v(x, a) = -v(x, -a), \\ &v(x, 0) \equiv 0, \qquad v(x, \pi) \equiv \pi. \end{aligned} \tag{4.4}$$

From (4.4), we shall consider $v(x, a)$ only for $0 < a < \pi$.

LEMMA 4.1. *Let $0 < a < \pi$. Then*

(i) $-\pi/2 < v(x, a) < \pi/2$ *for* $0 < a < \pi/2$, $x \geqq 0$.

(ii) $-\pi < v(x, a) < \pi$ *for* $\pi/2 \leqq a < \pi$, $x \geqq 0$.

(iii) $v(x, a)$ *is oscillatory over* $[0, \infty)$ *for all* $0 < a < \pi$.

Proof. Multiplying (4.3) by $v'(x)$ and integrating the resulting equation from 0 to x, we obtain

$$\frac{1}{2}(v'(x))^2 = x \cos v(x) - \int_0^x \cos v(\xi)\, d\xi \geqq 0. \tag{4.5}$$

If $0 < a < \pi/2$, then $\cos a = \cos v(0) > 0$. We claim that $\cos v(x) > 0$ for all $x \geqq 0$. If not, then there exists $x_0 > 0$ such that $\cos v(x) > 0$ for all $0 \leqq x < x_0$ and $\cos v(x_0) = 0$. Then this contradicts (4.5) with $x = x_0$ and we complete the proof for (i).

If $\pi/2 \leqq a < \pi$, then $\cos a = \cos v(0) \in (-1, 0]$. We claim that $\cos v(x) \neq -1$ for all $x \geqq 0$. If not, then there exists $x_0 > 0$ such that $\cos v(x_0) = -1$ and $\cos v(x) > 1$ for $0 \leqq x < x_0$. Again from (4.5) we obtain a contradiction. Hence $-\pi < v(x, a) < \pi$ for all $x \geqq 0$ and we established (ii).

We next show that $v(x, a)$ is oscillatory over $[0, \infty)$ for any $0 < a < \pi$. Let

$$V(x) = (1 - \cos v(x)) + \frac{1}{2}\frac{(v'(x))^2}{x}.$$

It is easy to verify that

$$V'(x) = -\frac{1}{2}\left(\frac{v'(x)}{x}\right)^2 \leqq 0.$$

Then we have

$$1 - \cos v(x) \leqq V(x) \leqq V(0) = 1 - \cos a.$$

Since $-\pi < v(x) < \pi$, we then have $|v(x)| \leqq a$ for all $x \geqq 0$. We rewrite the equation in (4.3) as

$$v''(x) + x\left(\frac{\sin v(x)}{v(x)}\right) v(x) = 0. \tag{4.6}$$

Let $0 < \delta < \min_{0 \leqq v \leqq a} (\sin v / v)$. Using Sturm's comparison theorem [5], we compare (4.6) with

$$v'' + \delta v = 0, \tag{4.7}$$

which is oscillatory over $[0, \infty)$. Thus we complete the proof for (iii). □

Next we introduce the following notation:

$$\Delta(x, a) = \frac{dv}{da}(x, a), \qquad \phi(x) = \Delta(x, 0).$$

Differentiating (4.3) with respect to a yields

$$\Delta''(x) + x(\cos v(x, a))\Delta(x) = 0, \quad \Delta(0) = 1, \quad \Delta'(0) = 0. \tag{4.8}$$

Setting $a = 0$ in (4.8) yields

$$\phi''(x) + x\phi(x) = 0, \quad \phi(0) = 1, \quad \phi'(0) = 0. \tag{4.9}$$

The equation in (4.9) is the well-known Airy equation which is oscillatory over $[0, \infty)$. Let λ_n, γ_n be the nth zero of $\phi(x)$ and $\phi'(x)$, respectively, for $n = 1, 2, \cdots$. We note that

$$\begin{aligned} &\lambda_1 \approx 1.98635, \quad \lambda_2 \approx 3.82557, \quad \lambda_3 \approx 5.29566, \\ &\lambda_4 \approx 6.58432, \text{ etc. (See Fig. 2.)} \end{aligned} \tag{4.10}$$

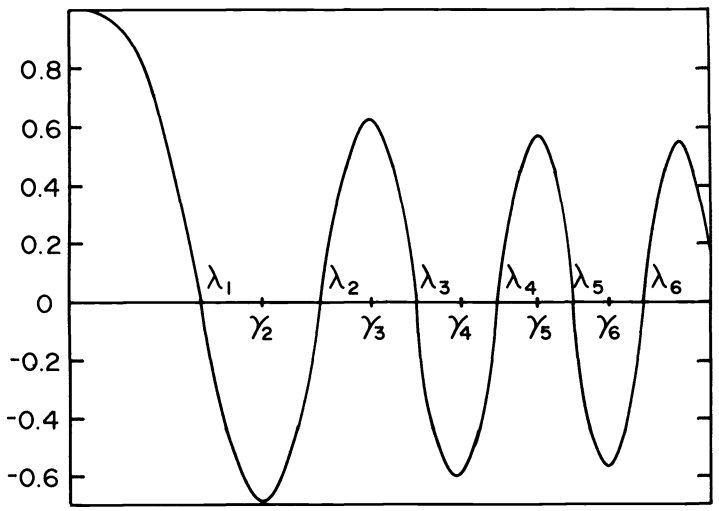


FIG. 2

From Lemma 4.1(iii), $v(x, a)$ is oscillatory over $[0, \infty)$ for any $0 < a < \pi$. Let $y_n(a)$, $z_n(a)$ be the nth zero of $v(x, a)$ and $v'(x, a)$, respectively, for $n = 1, 2, \cdots$, $0 < a < \pi$. (See Fig. 3.)

LEMMA 4.2.

(i) $\lim_{a \to 0^+} y_n(a) = \lambda_n$, $\lim_{a \to 0^+} z_n(a) = \gamma_n$ *for* $n = 1, 2, \cdots$,

(ii) $\lim_{a \to \pi^-} y_n(a) = +\infty$ *for* $n = 1, 2, \cdots$.

Proof. The proof of (i) follows directly from the following identities [6]:

$$\lim_{a \to 0^+} \frac{v(x, a)}{a} = \lim_{a \to 0^+} \frac{v(x, a) - v(x, 0)}{a}$$

$$= \lim_{a \to 0^+} \frac{dv}{da}(x, \delta_a), \qquad 0 < \delta_a < a$$

$$= \frac{dv}{da}(x, 0) = \phi(x),$$

and

$$\lim_{a \to 0^+} \frac{v'(x, a)}{a} = \lim_{a \to 0^+} \frac{v'(x, a) - v'(x, 0)}{a}$$

$$= \lim_{a \to 0^+} \frac{d}{da}(v'(x, \delta_a)) \qquad 0 < \delta_a < a$$

$$= \frac{d}{dx}\left(\frac{d}{da} v(x, 0)\right) = \phi'(x)$$

since $v(x, \pi) \equiv \pi$ and $v(x, a)$ are oscillatory over $[0, \infty)$ for $0 < a < \pi$. From continuous dependence on initial values, we obtain $\lim_{a \to \pi^-} y_1(a) = +\infty$ and hence $\lim_{a \to \pi^-} y_n(a) = +\infty$ for $n = 1, 2, \cdots$. Thus we complete the proof for (ii). □

In addition to the properties (i), (ii) in Lemma 4.2., we shall show that $y_n(a)$ satisfies

$$\frac{dy_n}{da} > 0 \quad \text{for all } n = 1, 2, \cdots \text{ and } 0 < a < \pi. \tag{4.11}$$

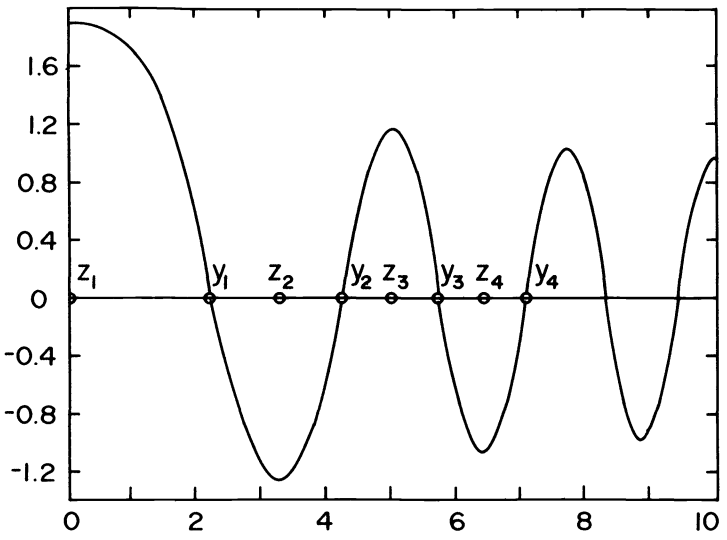


FIG. 3

Assume that (4.11) holds; then we may plot the following graphs for $y_n(a)$, $n = 1,2,\cdots$. (See Fig. 4.) Then we conclude from (4.2) and (4.4) that

If $0 < K < \lambda_1$ then (4.2) has the unique solution $v(x) \equiv 0$.

If $\lambda_1 < K < \lambda_2$ then (4.2) has three distinct solutions.

If $\lambda_n < K < \lambda_{n+1}$ then (4.2) has $2n+1$ distinct solutions.

Since

$$v(y_n(a), a) = 0, \qquad 0 < a < \pi, \tag{4.12}$$

differentiating (4.12) with respect to a yields

$$v'(y_n(a), a)\frac{dy_n}{da} + \frac{dv}{da}(y_n(a), a) = 0,$$

or

$$\frac{dy_n}{da} = -\frac{\Delta(y_n(a), a)}{y'(y_n(a), a)}. \tag{4.13}$$

We now state our main result.

THEOREM 1. *Let* $0 < a < \pi$.

(i) *The solution* $v(x, a)$ *of* (4.3) *has an infinite number of isolated zeros* $y_n(a)$, $y_1 < y_2 < \cdots < y_n$ *and* $y_n \to \infty$ *as* $n \to \infty$; *likewise* $v'(x, a)$ *has an infinite number of isolated zeros,* $z_n(a)$, $0 = z_1 < z_2 < \cdots < z_n$, *interlacing the* y_n; *furthermore*

$$\lim_{a \to 0^+} y_n(a) = \lambda_n, \qquad \lim_{a \to 0^+} z_n(a) = \gamma_n,$$

and

$$\lim_{a \to \pi^-} y_n(a) = \infty \qquad \text{for } n = 1, 2, \cdots.$$

(ii) $y_n(a)$ *is a differentiable function of* a *and*

$$\frac{dy_n}{da} > 0 \quad \text{for } n = 1, 2, \cdots.$$

We have shown part (i) in the above lemmas. The proof of (ii) follows directly from (4.13) and Lemma 4.3 below.

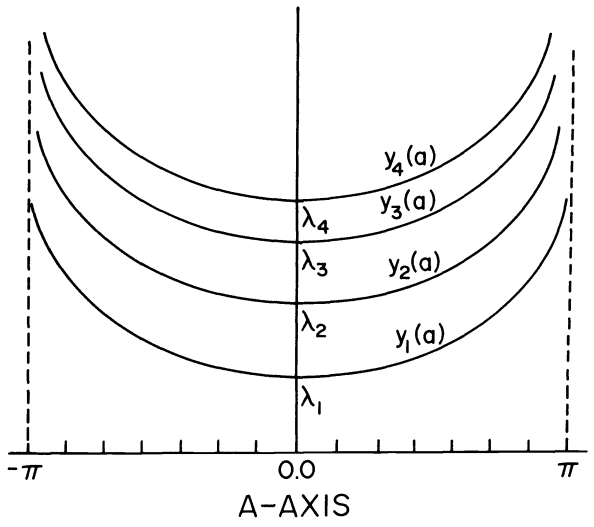


FIG. 4

LEMMA 4.3. *Let* $0<a<\pi$. *Then* $\Delta(x, a)$ *has an infinite number of isolated zeros* $\alpha_n(a)$, $0<\alpha_1<\cdots<\alpha_n$. $\Delta'(x, a)$ *satisfies the following*:

(i) *If* $0<a<\pi/2$, *then* $\Delta'(x, a)$ *has an infinite number of isolated zeros* $\beta_n(a)$, $0=\beta_1<\beta_2<\cdots<\beta_n$. *Furthermore* $\beta_1=z_1=0<y_1<\alpha_1<z_2<\beta_2<y_2<\alpha_2<\cdots<y_n<\alpha_n<z_{n+1}<\beta_{n+1}<y_{n+1}$. (See Fig. 5.)

(ii) *If* $\pi/2\leqq a<\pi$ *then* $\Delta'(x, a)$ *has an infinite number of isolated zeros* $\beta_n(a)$, $0=\beta_0<\beta_1<\cdots<\beta_n$. *Furthermore* $\beta_0=z_1=0<\beta_1<y_1<\alpha_1<z_2<\beta_2<y_2<\cdots<y_n<\alpha_n<z_{n+1}<\beta_{n+1}<y_{n+1}$. (See Fig. 6.)

Before we prove Lemma 4.3 we consider (4.3) and (4.8). Let

(A) $$v''+x\sin v=0,\quad v(0)=a,\quad v'(0)=0,$$

(B) $$\Delta''+x(\cos v)\Delta=0,\quad \Delta(0)=1,\quad \Delta'(0)=0.$$

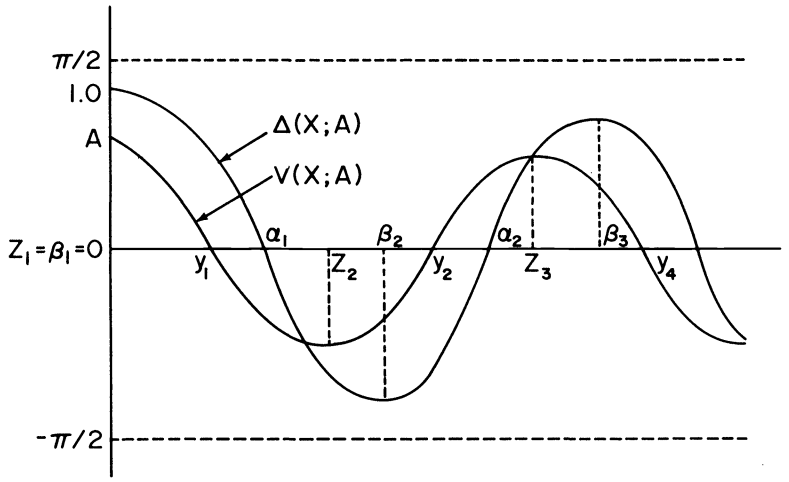


FIG. 5

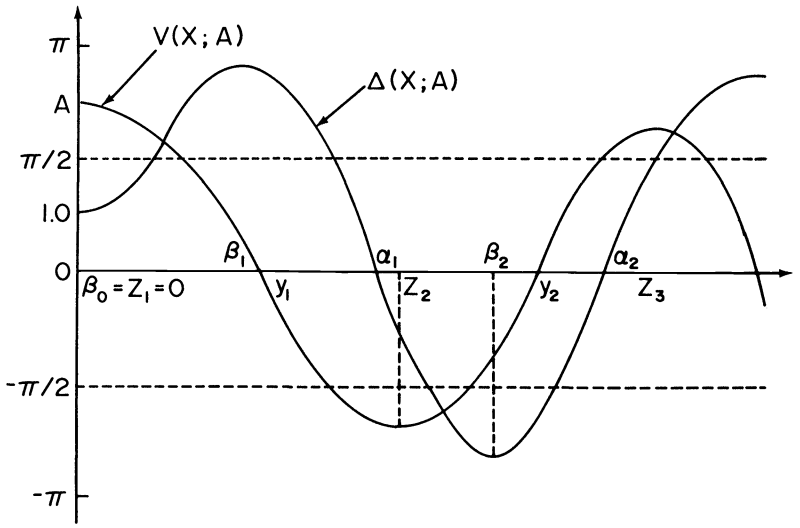


FIG. 6

In addition to (A) and (B), we form the following equations satisfied by Δ' and $\hat{v} = (x - 3y_n)v'$, respectively:

(C) $$(\Delta')'' + x \cos v\Delta' = \Delta(xv' \sin v - \cos v),$$

(D) $$\hat{v}'' + x \cos v\hat{v} = -3(x - y_n) \sin v.$$

Multiplying (A) by Δ and multiplying (B) by v, subtracting the resulting equations from each other and integrating the final expression from α to β, we obtain

(a) $$(v'\Delta - v\Delta')\Big|_\alpha^\beta = \int_\alpha^\beta x\Delta v\left(\cos v - \frac{\sin v}{v}\right) dx.$$

Multiplying (A) by Δ' and multiplying (C) by v, subtracting the resulting equations from each other and integrating the final expression from α to β, we obtain

(b) $$(v'\Delta' - v\Delta'')\Big|_\alpha^\beta = \int_\alpha^\beta \{-x\Delta' \sin v + x\Delta' v \cos v + \Delta v(\cos v - xv' \sin v)\}\, dx.$$

Multiplying (D) by Δ and multiplying (B) by $\hat{v}$, subtracting the resulting equation from each other and integrating the final expression from α to β, we obtain

(c) $$(\hat{v}'\Delta - \hat{v}\Delta')\Big|_\beta^\alpha = -\int_\beta^\alpha 3(x - y_n)\Delta \sin v\, dx.$$

Finally we observe that, since $v'(0) = 0$, $v(0) = a$, $\Delta(0) = 1$, $\Delta'(0) = 0$, $0 < a < \pi$,

$$\operatorname{sg} v = (-1)^n \quad \text{for } y_n < x < y_{n+1},$$
$$\operatorname{sg} v' = (-1)^n \quad \text{for } z_n < x < z_{n+1},$$
$$\operatorname{sg} \Delta = (-1)^n \quad \text{for } \alpha_n < x < \alpha_{n+1},$$
$$\operatorname{sg} \Delta' = (-1)^n \quad \text{for } \beta_n < x < \beta_{n+1}.$$

Proof of Lemma 4.3. We shall prove the lemma by induction. We assume the truth of the statement up to α_m, for $m = 1$.

If $\pi/2 \leqq a < \pi$, then we claim that $\beta_1 < y_1$. If not, $\beta_1 \geqq y_1$, then $\Delta'(x) > 0$ for all $0 \leqq x \leqq y_1$. We specialize (α, β) in (b) to $(0, y_1)$. Then we obtain

$$\begin{aligned} v'\Delta' - v\Delta''\Big|_0^{y_1} &= \int_0^{y_1} (v\Delta' x \cos v - x\Delta' \sin v + \Delta v \cos v)\, dx \\ &\quad + \int_0^{y_1} x\Delta v(-v' \sin v)\, dx. \end{aligned} \tag{4.14}$$

Since

$$\int_0^{y_1} x\Delta v(-v' \sin v)\, dx = x\Delta v \cos v\Big|_0^{y_1} - \int_0^{y_1} \cos v(x\Delta' v + x\Delta v' + \Delta v)\, dx$$

and $v'(0) = 0$, $\Delta''(0) = 0$, $v(y_1) = 0$. Then (4.14) becomes

$$\begin{aligned} v'(y_1)\Delta'(y_1) &= \int_0^{y_1} -x\Delta' \sin v\, dx - \int_0^{y_1} x\Delta v' \cos v\, dx \\ &= \int_0^{y_1} \Delta \sin v\, dx. \end{aligned} \tag{4.15}$$

This is a desired contradiction for $v'(y_1) < 0$, $\Delta'(y_1) > 0$ and $\Delta(x) > 0$, $\sin v(x) > 0$ for $0 \leqq x < y_1$.

We shall now show that $\alpha_1 > y_1$ for $0 < a < \pi$. If not, then there exists $\alpha^* \in (0, y_1)$ such that $\Delta(\alpha^*) = 0$, $\Delta'(\alpha^*) < 0$ and $\Delta(x) > 0$ for $0 \leqq x < \alpha^*$. We specialize (α, β) in (a) to $(0, \alpha^*)$. Then we have

$$-v(\alpha^*)\Delta'(\alpha^*) = \int_0^{\alpha^*} x\Delta v\left(\cos v - \frac{\sin v}{v}\right) dx. \tag{4.16}$$

Since $\cos v \leqq \sin v/v$ for $-\pi < v < \pi$ and $\Delta(x) > 0$, $v(x) > 0$ for $0 \leqq x < \alpha^*$, it follows that the right-hand side of (4.16) is negative. However, the left-hand side of (4.16) is positive. This leads to a contradiction.

We now want to complete the induction. For $0 < a < \pi$, we want to show the following:

(i) $y_m < \alpha_m < z_{m+1}$. By induction hypothesis $y_m < \alpha_m$. We want to show that $\alpha_m < z_{m+1}$. For $a = 0$, it is obvious that $\alpha_m(0) = \lambda_m$. From Lemma 4.2, we have

$$\lim_{a \to 0^+} z_{m+1}(a) = \gamma_{m+1} > \lambda_m.$$

By continuous dependence on parameter a, we have that $\alpha_m(a) < z_{m+1}(a)$ for $a > 0$ sufficiently small. We claim that $\alpha_m(a) < z_{m+1}(a)$ for all $0 < a < \pi$. If not, there exists $a^* \in (0, \pi)$ such that $\alpha_m(a^*) = z_{m+1}(a^*)$. We now specialize (α, β) in (c) to $(z_m(a^*), z_{m+1}(a^*))$ and $n = m$. Then we obtain

$$\hat{v}'\Delta - \hat{v}\Delta'\Big|_{z_m}^{z_{m+1}} = \int_{z_m}^{z_{m+1}} 3(x - y_m)(-\sin v(x))\Delta(x)\, dx \tag{4.17}$$

since $\hat{v}(z_{m+1}) = \hat{v}(z_m) = 0$, $z_{m+1} = \alpha_m$, $\hat{v}'(z_m) = (z_m - 3y_m)v''(z_m)$. Then (4.17) becomes

$$0 = (z_m - 3y_m)v''(z_m)\Delta(z_m) + \int_{z_m}^{z_{m+1}} 3(x - y_m)(-\sin v(x))\Delta(x)\, dx. \tag{4.18}$$

It is easy to verify that the right-hand side of (4.18) is positive. Then this is a desired contradiction. Hence, we have $\alpha_m(a) < z_{m+1}(a)$ for all $0 < a < \pi$.

(ii) $z_{m+1} < \beta_{m+1} < y_{m+1} < \alpha_{m+1}$. First we show that $z_{m+1} < \beta_{m+1}$. If not, then $\alpha_m < \beta_{m+1} \leqq z_{m+1}$. We specialize (α, β) in (a) to (α_m, β_{m+1}). Then we obtain

$$v'(\beta_{m+1})\Delta(\beta_{m+1}) + v(\alpha_m)\Delta'(\alpha_m) = \int_{\alpha_m}^{\beta_{m+1}} x\Delta v\left(\cos v - \frac{\sin v}{v}\right) dx. \tag{4.19}$$

It is easy to verify that the left-hand side of (4.19) is positive while the right-hand side is negative. This is a contradiction.

Next we show that $\beta_{m+1} < y_{m+1}$. If not, then $\beta_{m+1} \geqq y_{m+1}$. We specialize (α, β) in (b) to (z_{m+1}, y_{m+1}). Following similar arguments in the case $\beta_1 < y_1$, we deduce that

$$v'(y_{m+1})\Delta'(y_{m+1}) = z_{m+1}\Delta(z_{m+1}) \sin v(z_{m+1}) + \int_{z_{m+1}}^{y_{m+1}} \Delta \sin v\, dx.$$

It is easy to verify $v'(y_{m+1})\Delta'(y_{m+1}) \leqq 0$, $z_{m+1}\Delta(z_{m+1}) \sin v(z_{m+1}) > 0$, and $\int_{z_{m+1}}^{y_{m+1}} \Delta \sin v\, dx > 0$. Thus we obtain a contradiction.

Finally, we want to show that $y_{m+1} < \alpha_{m+1}$. If not, then $y_{m+1} \geqq \alpha_{m+1}$. We specialize (α, β) in (a) to $(\beta_{m+1}, \alpha_{m+1})$. Then we have

$$-[v'(\beta_{m+1})\Delta(\beta_{m+1}) + v(\alpha_{m+1})\Delta'(\alpha_{m+1})] = \int_{\beta_{m+1}}^{\alpha_{m+1}} x\Delta v\left(\cos v - \frac{\sin v}{v}\right) dx. \tag{4.20}$$

It is easy to verify that the left-hand side of (4.20) is positive while the right-hand side is negative. This is a contradiction. □

5. Numerical studies for $\alpha \neq \pi$ and discussions. In this section, we present our numerical studies for the multiplicities of the solutions of the problem $(P)_\alpha$, $0 < \alpha < \pi$. The analytic results in § 4 shall confirm that our numerical results are reliable.

Consider our bifurcation problem

$$(P)_\alpha \qquad \frac{d^2\psi}{ds^2} = K^2 s \sin\psi, \qquad K > 0,$$
$$\psi'(0) = 0, \quad \psi(1) = \alpha, \quad 0 < \alpha < \pi$$

and its scaled form

$$\frac{d^2\Psi}{ds^2} = s \sin\Psi, \quad \Psi'(0) = 0, \quad \Psi(K) = \alpha. \tag{2.6}$$

Let $\Psi(s, a)$ be the solution of the following initial value problem:

$$\frac{d^2\Psi}{ds^2} = s \sin\Psi, \quad \Psi'(0) = 0, \quad \Psi(0) = a. \tag{5.1}$$

It is easy to verify the following relations:

$$\Psi(s, a + 2\pi) = \Psi(s, a) + 2\pi, \tag{5.2}$$

$$\Psi(s, 2\pi - a) = 2\pi - \Psi(s, a). \tag{5.3}$$

For any $K > 0$, we consider the map

$$a \to \Psi(K, a), \qquad 0 \leqq a \leqq 2\pi.$$

Since $0 < \alpha < \pi$, from (5.3) we only need to compute numerically for $0 < a < \pi$. In the following, we used the ODE Solver DGEAR of the IMSL Library to compute the function $\alpha = \Psi(K, \alpha)$, $0 < a < \pi$, for various K.

In Fig. 7, the parameter K satisfies $0 < K = 1.0 < \lambda_1 \approx 1.98635$ and the graph $\alpha = \Psi(K, a)$ intersects $\alpha = \pi$ at only one point. *We conjecture that for* $0 < K < \lambda_1$ *the problem* $(P)_\alpha$ *has a unique solution for every* $\alpha \in (0, \pi)$. In Fig. 8, the parameter K satisfies

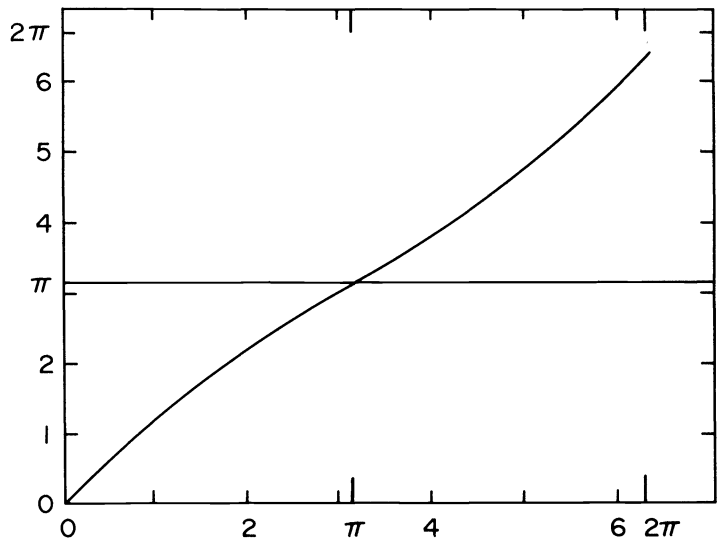


FIG. 7

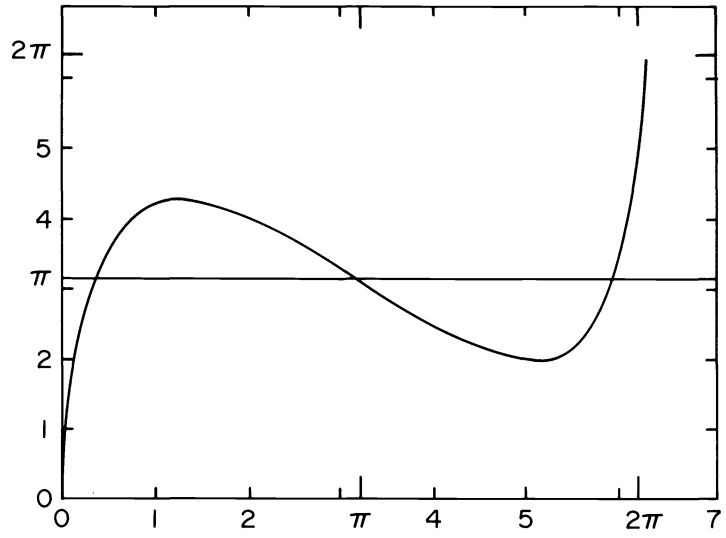


FIG. 8

$\lambda_1 < K = 3.0 < \lambda_2 \approx 3.82557$ and the graph $\alpha = \Psi(K, a)$ intersects $\alpha = \pi$ at three distinct points. It shows that if $\lambda_1 < K < \lambda_2$, then the problem $(P)_\alpha$ has at most three distinct solutions. In Fig. 9, the parameter K satisfies $\lambda_2 < K = 4.5 < \lambda_3 \approx 5.29566$ and the graph $\alpha = \Psi(K, a)$ intersects $\alpha = \pi$ at five distinct points. It shows that if $\lambda_2 < K < \lambda_3$, then the problem $(P)_\alpha$ has at most five distinct solutions. *We conjecture that for* $\lambda_n < K < \lambda_{n+1}$ *the problem* $(P)_\alpha$ *has* $1, 3, \cdots, 2n+1$ *solutions for various* α.

Acknowledgments. We thank Professor C. Y. Wang of Michigan State University for suggesting this problem to us and for stimulating discussions. The authors also thank a referee for his comments.

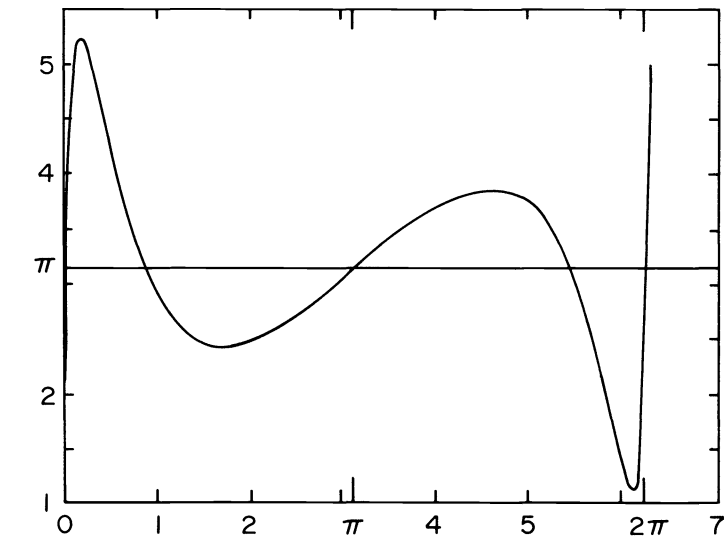


FIG. 9

REFERENCES

[1] W. G. BICKELEY, *The heavy elastica*, Phil. Mag. Ser., 717 (1934), pp. 603-622.

[2] S. N. CHOW, J. MALLET-PARET, AND J. YORKE, *Finding zeros of maps: homotopy methods that are constructive with probability one*, Math. Comp., 32 (1978), pp. 887-899.

[3] L. EULER, *De curvis elasticis*, 1744.

[4] A. G. GREENHILL, *Determination of the greatest height consistent with stability that a vertical pole or mast can be made, and of the greatest height to which a tree of given proportions can grow*, Proc. Cambridge Philos. Soc., 4 (1881), pp. 66-78.

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

[6] I. I. KOLODNER, *Heavy rotating string—A nonlinear eigenvalue problem*, Comm. Pure Appl. Math., 8 (1955), pp. 395-408.

[7] W. M. NI, *Uniqueness of solutions of nonlinear Dirichlet problems*, J. Differential Equations, 50 (1983), pp. 289-304.

[8] C. Y. WANG, *Large deformation of a heavy cantilever*, Quart. Appl. Math., 39 (1981), pp. 261-273.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

008

ON THE UNIQUENESS OF A LIMIT CYCLE FOR A PREDATOR-PREY SYSTEM*

LII-PERNG LIOU† AND KUO-SHUNG CHENG‡

Abstract. The uniqueness of a limit cycle for a predator-prey system is proved in this paper. The method used is an improvement of the method used earlier by Cheng.

Key words. limit cycle, preditor-prey system

AMS(MOS) subject classifications. primary 34D05; secondary 34C15

1. Introduction. Stability analysis for a nontrivial periodic solution of ordinary differential equations is very rare and difficult to obtain even in a two-dimensional system. One well-known example is the Lienard equation, in particular, the Van der Pol equation. See Hartman [6] and Hirsch and Smale [7] for details. For biological predator-prey systems, Hsu, Hubbell, and Waltman [8], [9] considered the following competing-predators system:

$$
\begin{aligned}
&\dot{S}(t) = rS(t)\left(1-\frac{S(t)}{K}\right)-\left(\frac{m_1}{y_1}\right)\left(\frac{X_1(t)S(t)}{a_1+S(t)}\right)-\left(\frac{m_2}{y_2}\right)\left(\frac{X_2(t)S(t)}{a_2+S(t)}\right),\\
&\dot{X}_1(t) = X_1(t)\left(\frac{m_1S(t)}{a_1+S(t)}-D_1\right), \qquad (1)\\
&\dot{X}_2(t) = X_2(t)\left(\frac{m_2S(t)}{a_2+S(t)}-D_2\right),\\
&S(0)=S_0>0, \quad X_i(0)=X_{i0}>0, \quad i=1,2,
\end{aligned}
$$

where $X_i(t)$ is the population of the ith predator at time t; $S(t)$ is the population of the prey at time t; m_i is the maximum growth rate of the ith predator; D_i is the death rate of the ith predator; y_i is the yield factor of the ith predator feeding on the prey; and a_i is the half-saturation constant of the ith predator, which is the prey density at which the functional response of the predator is half maximal. The parameters r and K are the intrinsic rate of increase and the carrying capacity for the prey population, respectively. Hsu, Hubbell, and Waltman analyzed solutions of this system and found that the behavior of solutions depends mainly on the two-dimensional system:

$$
\begin{aligned}
&\dot{S}(t) = rS(t)\left(1-\frac{S(t)}{K}\right)-\left(\frac{m}{y}\right)\left(\frac{x(t)S(t)}{a+S(t)}\right),\\
&\dot{x}(t) = x(t)\left(\frac{mS(t)}{a+S(t)}-D_0\right), \qquad (2)\\
&S(0)=S_0>0, \qquad x(0)=x_0>0,
\end{aligned}
$$

* Received by the editors January 15, 1986; accepted for publication (in revised form) May 11, 1987. This work was supported in part by the National Science Council of the Republic of China.

† Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan, Republic of China.

‡ Institute of Applied Mathematics, National Tsing Hua University, Hsinchu, Taiwan 300, Republic of China.

where r, K, m, y, a, and D_0 are positive constants. They analyzed system (2) and found that if $\lambda < (K-a)/2$, where $\lambda = a/(b-1)$ and $b = m/D_0$, then the unique interior equilibrium point (λ, x^*) is unstable. They conjectured that the system (2) has a unique stable limit cycle in this case. This conjecture was answered affirmatively by Cheng in [2]. In these examples, symmetric properties are an important ingredient of the proof. The Van der Pol equations are

$$\begin{aligned} \dot{x} &= y - (x^3 - x), \\ \dot{y} &= -x. \end{aligned} \tag{3}$$

The isocline $\dot{x} = 0$, i.e., the curve $y = x^3 - x$, is symmetric with respect to the origin. This fact is important in the analysis of (3). For the system (2), the isocline $\dot{S} = 0$ is the curve

$$x = r(y/m)(1 - S/K)(a + S). \tag{4}$$

This curve is part of a parabola and hence is also symmetric with respect to the line $S = (K-a)/2$. The proof of Cheng [2] uses this symmetry property in an essential way. From the point of view of perturbation theory, there is no reason to believe that some symmetry properties are indispensable for a stable limit cycle. In this respect, if we can devise a proof that is valid for a more general "nonsymmetric" system, even if it is only a slight generalization, we will feel comfortable with it.

The purpose of this paper is to improve our method used in [2] to prove the uniqueness of a limit cycle for a more general predator-prey system without the symmetry properties of the isocline. At the end of our proof, we also close a gap in the original proof given in [2].

2. The equations and statements of the main result. We will consider the following predator-prey system:

$$\begin{aligned} &\dot{x} = x(f(x) - y), \\ &\dot{y} = y(g(x) - \lambda), \\ &x(0) = x_0 > 0, \qquad y(0) = y_0 > 0, \qquad \lambda > 0. \end{aligned} \tag{5}$$

Note that if $g(x) = x$ and $f(x) = (1 - x/K)(a + x)$, then the system (5) is essentially equivalent to the system (2) up to some irrelevant constants. Our general assumptions about $f(x)$ and $g(x)$ are:

(i) $g \in C^1([0, \infty))$, $g(0) = 0$, $g'(x) > 0$ for all $x \geqq 0$.

(ii) $f \in C^2([0, \infty))$, $f(0) \geqq 0$, and there exists $K > 0$ such that $f(K) = 0$ and $(x - K)f(x) < 0$ for $x \neq K$. There exists an a, $0 < a < K$, such that $f'(x) > 0$ for $0 < x < a$, $f'(a) = 0$ and $f'(x) < 0$ for $a < x$.

(iii) $g(x^*) = \lambda$, $y^* = f(x^*)$, and $0 < x^* < a$.

(iv) $(d/dx)(xf'(x))/(g(x) - \lambda)) < 0$ for $x < x^*$ and $x > \bar{x}^*$, where $\bar{x}^* = f_2^{-1} \circ f_1(x^*)$ and $f_1 = f|_{(0,a)}$, $f_2 = f|_{(a,K)}$.

The phase plane of (5) under assumptions (i)-(iv) is roughly as shown in Fig. 1. We consider only the case $x^* < a$. In the case $a < x^* < K$, the equilibrium point (x^*, y^*) is locally asymptotically stable. We refer to Cheng, Hsu, and Lin [3] for global stability analysis.

Note that if

$$\begin{aligned} g(x) &= x, \\ f(x) &= F(x) + \varepsilon H(x), \end{aligned}$$

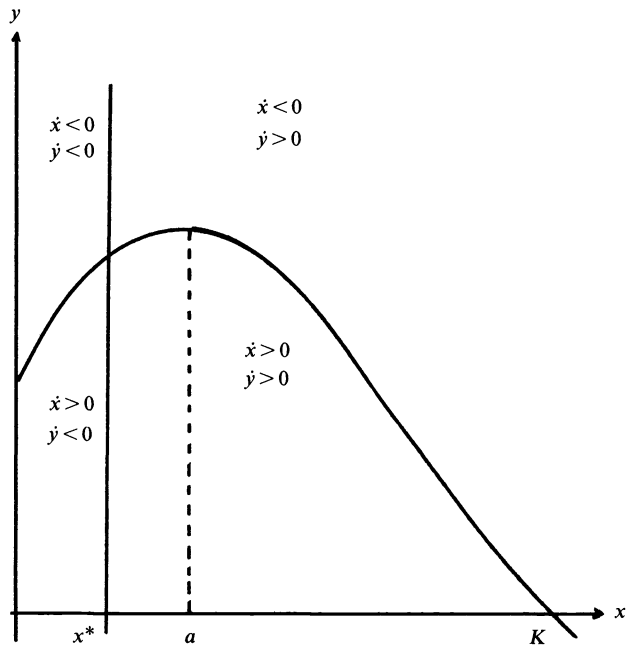


FIG. 1

then g and f satisfy assumptions (i)-(iv) if $\varepsilon > 0$ is sufficiently small where

$$F(x) = (1-x)(b+x)$$

and $H(x)$ is a C^2 function satisfying

$$H'(x) \geqq 0 \quad \text{for } 0 < x < a,$$

$$H'(x) \leqq 0 \quad \text{for } a < x,\ a = \frac{1-b}{2}.$$

In fact,

$$\frac{d}{dx}\left(\frac{xf'(x)}{x-\lambda}\right) = \frac{-1}{(x-\lambda)^2}\{[2(x-\lambda)^2 + 2\lambda(a-\lambda)] - \varepsilon[(x-\lambda)xH''(x) - \lambda H'(x)]\}.$$

Thus if $(a-\lambda)$ is reasonably large, we can allow ε to be reasonably large and the isocline

$$y = F(x) + \varepsilon H(x)$$

can be quite unsymmetric with respect to the line $x = a$.

Our main result follows.

THEOREM 1. *Under the assumptions* (i)-(iv), (5) *posseses a unique limit cycle which is globally stable.*

3. Proof of Theorem 1. We need some lemmas.

LEMMA 1. *The solutions* $x(t)$, $y(t)$ *of* (5) *are positive and bounded.*

LEMMA 2. *The unique interior equilibrium point* (x^*, y^*) *of* (5) *is a source.*

LEMMA 3. *Let* Γ *be a nontrivial closed orbit of* (2). *Then*

$$\Gamma \subset \{(x, y): 0 < x < K, 0 < y\}.$$

Let L, R, H, and J be the leftmost, rightmost, highest, and lowest points of Γ, *respectively. Then*

$$L \in \{(x, y): 0 < x < x^*, y = f(x)\},$$

$$R \in \{(x, y): x^* < x < K, y = f(x)\},$$

$$H \in \{(x, y): x = x^*, y > y^*\},$$

$$J \in \{(x, y): x = x^*, 0 < y < y^*\}.$$

The proof of Lemma 1 is given in Albrecht et al. [1]. Lemma 2 follows from a straightforward calculation and Lemma 3 is easy enough. Hence we omit all the proofs of these lemmas.

Before we state and prove our next lemma, we define a transformation T from $(0, a) \times (0, \infty)$ to $(a, K) \times (0, \infty)$,

$$\begin{aligned} T(x, y) &\equiv (T_1(x, y), T_2(x, y)) \\ &\equiv (f_2^{-1} \circ f_1(x), y), \end{aligned} \tag{6}$$

where f_1 and f_2 are the restriction of f on $(0, a)$ and (a, K), respectively. From assumption (ii), it is easy to see that T is a one-to-one transformation.

Now, we can state our main lemmas.

LEMMA 4. *Let* Γ *be a nontrivial closed orbit of* (5). Γ *meets the vertical line* $x = a$ *at points A and B with y-coordinates* $y_B > y_A$. (*See Fig.* 2.) *Let the image of* arc $\widehat{BHLJA}$ *of* Γ *under the transformation T be* $\widehat{BH'L'J'A}$. *Then* arc $\widehat{H'L'J'}$ *intersects* arc $\widehat{BRA}$ *of* Γ *at exactly two points* $P = (x_1, y_1)$ *and* $Q = (x_2, y_2)$ *with* $y_1 > f(x_1)$ *and* $y_2 < f(x_2)$.

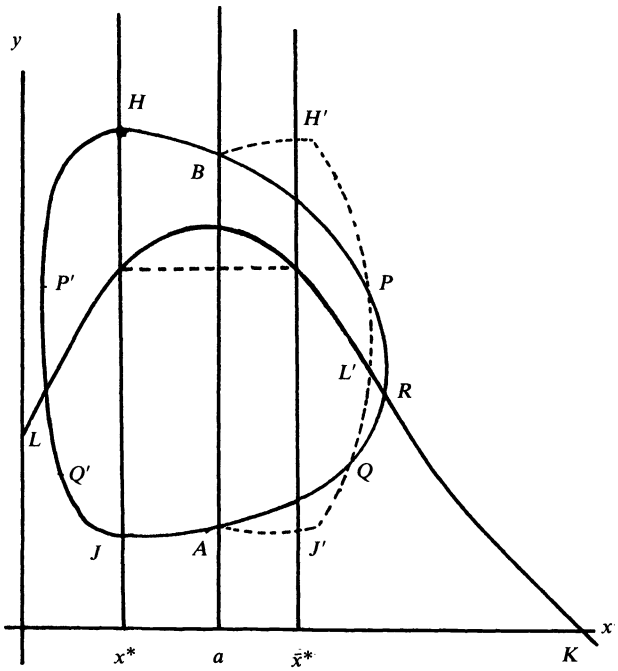


FIG. 2

Furthermore, let $P' = (x_1', y_1') = T^{-1}(P)$ *and* $Q' = (x_2', y_2') = T^{-1}(Q)$. *Then*

$$0 > \frac{x_1' f_1'(x_1')}{g(x_1') - \lambda} \geqq \frac{x_1 f_2'(x_1)}{g(x_1) - \lambda}, \tag{7}$$

$$0 > \frac{x_2' f_1'(x_2')}{g(x_2') - \lambda} \geqq \frac{x_2 f_2'(x_2)}{g(x_2) - \lambda}. \tag{8}$$

Proof. Consider the function

$$V(x, y) = \int_{x^*}^{x} \frac{g(\xi) - \lambda}{\xi}\, d\xi. \tag{9}$$

Then

$$\frac{dV(x(t), y(t))}{dt} = [g(x(t)) - \lambda][f(x(t)) - y(t)]. \tag{10}$$

Let the period of Γ be τ. We have

$$\int_0^{\tau} \frac{dV(x(t), y(t))}{dt}\, dt = 0. \tag{11}$$

On the other hand, we have

$$\begin{aligned} \int_0^{\tau} \frac{dV(x(t), y(t))}{dt}\, dt &= \int_0^{\tau} [g(x(t)) - \lambda][f(x(t)) - y(t)]\, dt \\ &= \oint_{\Gamma} (f(x) - y)\frac{dy}{y}. \end{aligned} \tag{12}$$

Let Ω_1 be the interior of the domain bounded by arc $\widehat{BHLJA}$ and line $x = a$ and Ω_2 be the interior of the domain bounded by arc $\widehat{BRA}$ and the line $x = a$. Also, let $\Omega = \Omega_1 \cup \Omega_2$ and $\Omega_1' = T(\Omega_1)$. From the definition of T it is easy to see arc $\widehat{BH'}$ lies above Γ and arc $\widehat{J'A}$ lies below Γ. Hence either

$$\widehat{BH'L'J'A} \cap \Omega_2 = \varnothing(\text{empty}) \quad \text{and} \quad \Omega_2 \subset \Omega_1' \tag{13}$$

or

$$S \equiv \widehat{BH'L'J'A} \cap \Omega_2 \neq \varnothing.$$

We now show that the assumption $\Omega_2 \subset \Omega_1'$ leads to a contradiction. From (11) and (12), we have

$$\begin{aligned} 0 &= \int_0^{\tau} \frac{dV(x(t), y(t))}{dt}\, dt \\ &= \oint_{\Gamma} \frac{1}{y}[f(x) - y]\, dy \\ &= \iint_{\Omega} \frac{f'(x)}{y}\, dx\, dy \quad \text{(Green's theorem)} \\ &= \iint_{\Omega_1} \frac{f'(x)}{y}\, dx\, dy + \iint_{\Omega_2} \frac{f'(x)}{y}\, dx\, dy \\ &= \iint_{\Omega_1} \frac{f_1'(x)}{y}\, dx\, dy + \iint_{\Omega_2} \frac{f_2'(x)}{y}\, dx\, dy. \end{aligned} \tag{14}$$

Now let $T: \Omega_1 \to \Omega_1'$ be the transformation defined in (6). Let

$$T(x, y) = (u, v).$$

Then

$$\begin{aligned} u &= f_2^{-1} \circ f_1(x), \\ v &= y. \end{aligned}$$

Hence $(x, y) = T^{-1}(u, v)$ and

$$\begin{aligned} x &= f_1^{-1} \circ f_2(u), \\ y &= v. \end{aligned} \tag{15}$$

The Jacobian of T^{-1} is

$$\begin{aligned} \frac{\partial(x, y)}{\partial(u, v)} &= \begin{vmatrix} (f_1^{-1})'(f_2(u)) \cdot f_2'(u) & 0 \\ 0 & 1 \end{vmatrix} \\ &= (f_1^{-1})'(f_2(u)) \cdot f_2'(u). \end{aligned} \tag{16}$$

But since $f_2'(u) < 0$ and $(f_1^{-1})'(f_2(u)) > 0$, we have

$$\left| \frac{\partial(x, y)}{\partial(u, v)} \right| = -(f_1^{-1})'(f_2(u)) \cdot f_2'(u). \tag{17}$$

Hence, we have from (17)

$$\begin{aligned} \iint_{\Omega_1} \frac{f_1'(x)}{y}\, dx\, dy &= \iint_{\Omega_1'} \frac{f_1'(f_1^{-1} \circ f_2(u))}{v} \cdot [-(f_1^{-1})'(f_2(u)) \cdot f_2'(u)\, du\, dv \\ &= -\iint_{\Omega_1'} \frac{f_1'(f_1^{-1} \circ f_2(u)) \cdot (f_1^{-1})'(f_2(u)) \cdot f_2'(u)}{v}\, du\, dv. \end{aligned} \tag{18}$$

But

$$\begin{aligned} f_1'(f_1^{-1} \circ f_2(u)) \cdot (f_1^{-1})'(f_2(u)) &= \frac{d}{dz}(f_1 \circ f_1^{-1}(z))\Big|_{z = f_2(u)} \\ &= 1. \end{aligned} \tag{19}$$

From (18) and (19), we obtain

$$\begin{aligned} \iint_{\Omega_1} \frac{f_1'(x)}{y}\, dx\, dy &= -\iint_{\Omega_1'} \frac{f_2'(u)}{v}\, du\, dv \\ &= -\iint_{\Omega_1'} \frac{f_2'(x)}{y}\, dx\, dy \quad \text{(identify } u = x, v = y\text{)}. \end{aligned} \tag{20}$$

Combining (13), (14), and (20), finally we have

$$\begin{aligned} 0 &= \int_0^{\tau} \frac{dV(x(t), y(t))}{dt}\, dt \\ &= \iint_{\Omega_1} \frac{f_1'(x)}{y}\, dx\, dy + \iint_{\Omega_2} \frac{f_2'(x)}{y}\, dx\, dy \\ &= -\iint_{\Omega_1'} \frac{f_2'(x)}{y}\, dx\, dy + \iint_{\Omega_2} \frac{f_2'(x)}{y}\, dx\, dy \\ &= -\iint_{\Omega_1' - \Omega_2} \frac{f_2'(x)}{y}\, dx\, dy \\ &> 0 \quad \text{(recall that } f_2'(x) < 0\text{)}. \end{aligned} \tag{21}$$

This is a contradiction. Hence $S = \widehat{BH'L'J'A} \cap \Omega_2 \neq \varnothing$.

Let $\bar{S}$ denote the closure of S and let $P=(x_1, y_1)$ and $Q=(x_2, y_2)$ be the "highest" and "lowest" points of $\bar{S}$, respectively. Then, P is the highest point and Q is the lowest point where arc $\widehat{BH'L'J'A}$ enters the region Ω_2 from the outside of Ω_2. It is easy to see that $y_1 > y_2$. First we assume that $y_1 > f(x_1)$. Let $(dy/dx)'_P$ and $(dy/dx)_P$ be the slopes of arcs $\widehat{BH'L'J'A}$ and $\widehat{BRA}$ at point P, respectively. Since arc $\widehat{BH'L'J'A}$ enters Ω_2 from the outside of Ω_2 at point P, we have

$$0 > \left(\frac{dy}{dx}\right)_P \geqq \left(\frac{dy}{dx}\right)'_P. \tag{22}$$

But we have

$$\begin{aligned}\left(\frac{dy}{dx}\right)_P &= \frac{y_1(g(x_1)-\lambda)}{x_1(f(x_1)-y_1)} \\ &= \frac{y_1(g(x_1)-\lambda)}{x_1(f_2(x_1)-y_1)},\end{aligned} \tag{23}$$

and

$$\begin{aligned}\left(\frac{dy}{dx}\right)'_P &= \left(\frac{dv}{du}\right)_{(u,v)=(x_1,y_1)} \\ &= \left(\frac{dy}{d(f_2^{-1}\circ f_1(x))}\right)_{(x,y)=(x'_1,y'_1)=T^{-1}(x_1,y_1)} \\ &= \frac{y'_1(g(x'_1)-\lambda)}{(f_2^{-1})'(f_1(x'_1))\cdot f'_1(x'_1)\cdot x'_1\cdot(f_1(x'_1)-y'_1)}.\end{aligned} \tag{24}$$

Since

$$f_1(x'_1) = f_1(f_1^{-1}\circ f_2(x_1)) = f_2(x_1), \tag{25}$$

$$y'_1 = y_1, \tag{26}$$

and

$$\begin{aligned}(f_2^{-1})'(f_1(x'_1))\cdot f'_2(f_2^{-1}\circ f_1(x'_1)) &= (f_2^{-1})'(f_1(x'_1))\cdot f'_2(x_1) \\ &= 1.\end{aligned} \tag{27}$$

We have from (24), (25), (26), and (27)

$$\left(\frac{dy}{dx}\right)'_P = \frac{y'_1(g(x'_1)-\lambda)}{x'_1(f_2(x_1)-y_1)\cdot(1/f'_2(x_1))\cdot f'_1(x'_1)}. \tag{28}$$

Thus from (22), (23), and (29) we obtain

$$0 > \frac{y_1(g(x_1)-\lambda)}{x_1(f_2(x_1)-y_1)} \geqq \frac{y_1(g(x'_1)-\lambda)}{x'_1\cdot f'_1(x'_1)\cdot(f_2(x_1)-y_1)\cdot(1/f'_2(x_1))}. \tag{29}$$

Finally we get

$$0 > \frac{x'_1 f'_1(x'_1)}{g(x'_1)-\lambda} \geqq \frac{x_1 f'_2(x_1)}{g(x_1)-\lambda}. \tag{30}$$

Now the arc $\widehat{PR}$ satisfies the following differential equations:

$$\left(\frac{dy}{dx}\right)_{\widehat{PR}} = \frac{y(g(x)-\lambda)}{x(f_2(x)-y)} \tag{31}$$

and the arc $\widehat{PL'}$ satisfies

$$\begin{aligned}\left(\frac{dy}{dx}\right)_{\widehat{PL'}} &= \left(\frac{dv}{du}\right)_{(u,v)=(x,y)} \\ &= \frac{y'(g(x')-\lambda)}{(f_2^{-1})'(f_1(x'))\cdot f_1'(x')\cdot x'\cdot(f_1(x')-y')} \\ &= \frac{y(g(x')-\lambda)}{x'(f_2(x)-y)\cdot(1/f_2'(x))\cdot f_1'(x')}\end{aligned} \tag{32}$$

as in (24)-(28).

From (31) and (32) we have

$$\left(\frac{dy}{dx}\right)_{\widehat{PR}} = \frac{g(x)-\lambda}{xf_2'(x)}\cdot\frac{yf_2'(x)}{f_2(x)-y}, \tag{33}$$

$$\left(\frac{dy}{dx}\right)_{\widehat{PL'}} = \frac{g(x')-\lambda}{x'f_1'(x')}\cdot\frac{yf_2'(x)}{f_2(x)-y}. \tag{34}$$

From the assumption (iv) and (30), we have

$$\frac{g(x')-\lambda}{x'f_1'(x')} < \frac{g(x_1')-\lambda}{x_1'f_1'(x_1')} \leqq \frac{g(x_1)-\lambda}{x_1f_2'(x_1)} < \frac{g(x)-\lambda}{xf_2'(x)} \tag{35}$$

for all $x_1 < x$ (hence $x' < x_1'$).

Hence we have

$$0 > \frac{g(x)-\lambda}{xf_2'(x)}\cdot\frac{yf_2'(x)}{f_2(x)-y} > \frac{g(x')-\lambda}{x'f_1'(x')}\cdot\frac{yf_2'(x)}{f_2(x)-y}$$

for all $x_1 < x$.

From a well-known comparison theorem we get

$$y(x)_{\widehat{PR}} > y(x)_{\widehat{PL'}} \quad \text{for } x_1 < x < x_{L'}, \tag{36}$$

where $x_{L'}$ is the x-coordinate of L'.

This proves that if $y_1 > f(x_1)$, then the arc $\widehat{BH'L'}$ intersects the arc $\widehat{BR}$ only at the point P.

Now assume that $y_2 < f(x_2)$. Let $(dy/dx)_Q'$ and $(dy/dx)_Q$ be the slopes of arcs $\widehat{BH'L'J'A}$ and $\widehat{BRA}$ at the point $Q = (x_2, y_2)$, respectively. Then since $y_2 < f(x_2)$, it is obvious that

$$0 < \left(\frac{dy}{dx}\right)_Q \leqq \left(\frac{dy}{dx}\right)_Q'. \tag{37}$$

By arguments similar to those in (23)-(28), we have

$$\begin{aligned}\left(\frac{dy}{dx}\right)_Q &= \frac{y_2(g(x_2)-\lambda)}{x_2(f_2(x_2)-y_2)} \\ &= \frac{g(x_2)-\lambda}{x_2f_2'(x_2)}\cdot\frac{y_2f_2'(x_2)}{f_2(x_2)-y_2},\end{aligned} \tag{38}$$

$$\left(\frac{dy}{dx}\right)_Q' = \frac{g(x_2')-\lambda}{x_2'f_1'(x_2')}\cdot\frac{y_2f_2'(x_2)}{f_2(x_2)-y_2}. \tag{39}$$

Hence from (37), (38), and (39) we obtain

$$0>\frac{x_2' f_1'(x_2')}{g(x_2')-\lambda}\geqq\frac{x_2 f_2'(x_2)}{g(x_2)-\lambda}. \tag{40}$$

By arguments similar to those in (31)-(36), we can prove that if $y_2<f(x_2)$, then the arc $\widehat{L'J'A}$ intersects the arc $\widehat{RA}$ only at the point Q. From the above conclusion, P cannot be one of the intersection points of arcs $\widehat{L'J'A}$ and $\widehat{RA}$. Hence we conclude that

$$y_1>f(x_1), \qquad y_2<f(x_2) \tag{41}$$

and P and Q are the only intersection points of arcs $\widehat{BH'L'J'A}$ and $\widehat{BRA}$. Hence (30) and (40) hold. This completes the proof of the lemma. □

LEMMA 5. *Let* Γ *be a nontrivial closed orbit of* (5) *as described in Lemma* 4. *Define*

$$h(x, y)=x(f(x)-y), \qquad k(x, y)=y(g(x)-\lambda).$$

Then

$$\oint_\Gamma \operatorname{Div}(h, k)\, dt\equiv\oint_\Gamma\left(\frac{\partial h(x, y)}{\partial x}+\frac{\partial k(x, y)}{\partial y}\right) dt<0. \tag{42}$$

Proof. From the definitions of h and k, we have

$$\frac{\partial h(x, y)}{\partial x}+\frac{\partial k(x, y)}{\partial y}=(f(x)-y)+(g(x)-\lambda)+xf'(x). \tag{43}$$

But since Γ is a closed orbit, we have

$$\oint_\Gamma [f(x)-y]\, dt=\oint_\Gamma \frac{\dot{x}}{x}\, dt=0$$

and

$$\oint_\Gamma [g(x)-\lambda]\, dt=\oint_\Gamma \frac{\dot{y}}{y}\, dt=0.$$

Thus

$$\oint_\Gamma \operatorname{Div}(h, k)\, dt=\oint_\Gamma xf'(x)\, dt. \tag{44}$$

We divide the integration along Γ into integration along several arcs, that is, we let

$$\oint_\Gamma=\int_{\widehat{AQ}}+\int_{\widehat{QRP}}+\int_{\widehat{PB}}+\int_{\widehat{BP'}}+\int_{\widehat{P'LQ'}}+\int_{\widehat{Q'A}}. \tag{45}$$

Consider the integration along $\widehat{Q'A}$ first. The arc $\widehat{Q'A}$ of Γ can be parametrized by $(x, y_1(x))$, where $x_2'\leqq x\leqq a$. Hence

$$\begin{aligned}\int_{\widehat{Q'A}} x(t)f'(x(t))\, dt&=\int_{x_2'}^{a}\frac{f'(x)}{f(x)-y_1(x)}\, dx\\&=\int_{x_2'}^{a}\frac{f_1'(x)}{f_1(x)-y_1(x)}\, dx.\end{aligned} \tag{46}$$

Now let

$$u=f_2^{-1}\circ f_1(x), \qquad x\in[x_2', a]$$

or

$$x = f_1^{-1} \circ f_2(u), \qquad u \in [a, x_2].$$

Then

$$\begin{aligned}\int_{x_2'}^{a} \frac{f_1'(x)}{f_1(x) - y_1(x)}\,dx &= \int_{x_2}^{a} \frac{f_1'(f_1^{-1} \circ f_2(u)) \cdot (f_1^{-1})'(f_2(u)) \cdot f_2'(u)}{f_1(f_1^{-1} \circ f_2(u)) - y_1(f_1^{-1} \circ f_2(u))}\,du \\ &= \int_{x_2}^{a} \frac{f_2'(u)}{f_2(u) - y_1(f_1^{-1} \circ f_2(u))}\,du \\ &= -\int_{a}^{x_2} \frac{f_2'(x)}{f_2(x) - y_1(f_1^{-1} \circ f_2(x))}\,dx. \end{aligned} \tag{47}$$

We parametrize the arc $\overset{\frown}{AQ}$ of Γ by $(x, y_2(x))$, where $x \in [a, x_2]$. Then

$$\begin{aligned}\int_{\overset{\frown}{AQ}} x(t) f'(x(t))\,dt &= \int_{a}^{x_2} \frac{f'(x)}{f(x) - y_2(x)}\,dx \\ &= \int_{a}^{x_2} \frac{f_2'(x)}{f_2(x) - y_2(x)}\,dx. \end{aligned} \tag{48}$$

Combining (47) and (48), we obtain

$$\begin{aligned}&\left(\int_{\overset{\frown}{Q'A}} + \int_{\overset{\frown}{AQ}}\right)(x(t) f'(x(t)))\,dt \\ &\quad = \int_{a}^{x_2} \frac{f_2'(x)[y_2(x) - y_1(f_1^{-1} \circ f_2(x))]\,dx}{[f_2(x) - y_2(x)][f_2(x) - y_1(f_1^{-1} \circ f_2(x))]}.\end{aligned}$$

But for $x \in (a, x_2)$, we have

$$\begin{aligned}&f_2'(x) < 0, \qquad y_2(x) - y_1(f_1^{-1} \circ f_2(x)) > 0, \\ &f_2(x) - y_2(x) > 0, \qquad f_2(x) - y_1(f_1^{-1} \circ f_2(x)) > 0.\end{aligned}$$

Hence

$$\left(\int_{\overset{\frown}{Q'A}} + \int_{\overset{\frown}{AQ}}\right)(x(t) f'(x(t)))\,dt < 0. \tag{49}$$

Next we parametrize arc $\overset{\frown}{BP'}$ of Γ by $(x, y_3(x))$ and arc $\overset{\frown}{PB}$ by $(x, y_4(x))$. Then

$$\begin{aligned}\int_{\overset{\frown}{BP'}} (x(t) f'(x(t)))\,dt &= \int_{a}^{x_1} \frac{f'(x)}{f(x) - y_3(x)}\,dx \\ &= \int_{a}^{x_1'} \frac{f_1'(x)}{f_1(x) - y_3(x)}\,dx. \end{aligned} \tag{50}$$

Let $x = f_1^{-1} \circ f_2(u)$ or $u = f_2^{-1} \circ f_1(x)$. Then from (50)

$$\begin{aligned}&\int_{\overset{\frown}{BP'}} (x(t) f'(x(t)))\,dt \\ &\quad = \int_{a}^{x_1} \frac{f_1'(f_1^{-1} \circ f_2(u)) \cdot (f_1^{-1})'(f_2(u)) f_2'(u)}{f_1(f_1^{-1} \circ f_2(u)) - y_3(f_1^{-1} \circ f_2(u))}\,du \\ &\quad = \int_{a}^{x_1} \frac{f_2'(u)}{f_2(u) - y_3(f_1^{-1} \circ f_2(u))}\,du \\ &\quad = \int_{a}^{x_1} \frac{f_2'(x)}{f_2(x) - y_3(f_1^{-1} \circ f_2(x))}\,dx.\end{aligned} \tag{51}$$

Now from the parametrization of arc $\widehat{PB}$, we have

$$\begin{aligned}\int_{\widehat{PB}}(x(t)f'(x(t)))\,dt &= \int_{x_1}^{a}\frac{f'(x)}{f(x)-y_4(x)}\,dx \\ &= -\int_{a}^{x_1}\frac{f_2'(x)}{f_2(x)-y_4(x)}\,dx.\end{aligned} \tag{52}$$

Combining (51) and (52) we obtain

$$\begin{aligned}&\left(\int_{\widehat{PB}}+\int_{\widehat{BP'}}\right)(x(t)f'(x(t)))\,dt \\ &\quad=\int_{a}^{x_1}\frac{f_2'(x)[y_3(f_1^{-1}\circ f_2(x))-y_4(x)]}{[f_2(x)-y_4(x)][f_2(x)-y_3(f_1^{-1}\circ f_2(x))]}\,dx \\ &\quad<0.\end{aligned} \tag{53}$$

Now let us assume that $x_1'\geqq x_2'$, i.e., $x_2\geqq x_1$. (The case $x_1'<x_2'$ can be treated in the same manner.) Let

$$L_1=\{(x,y)\colon x=x_1',\ y_2'\leqq y\leqq y_1'\},$$
$$L_2=\{(x,y)\colon y=y_2',\ x_2'\leqq x\leqq x_1'\}.$$

We parametrize the arc $\widehat{P'LQ'}$ by $(h_1(y),y)$ and let the domain bounded by the arc $\widehat{P'LQ'}$, L_2 and L_1 be denoted by D_1. Then we have

$$\begin{aligned}&\int_{\widehat{P'LQ}}(x(t)f'(x(t)))\,dt \\ &\quad=\int_{\widehat{P'LQ'}}\frac{xf'(x)}{y[g(x)-\lambda]}\bigg|_{x=h_1(y)}dy \\ &\quad=\left(\int_{\widehat{P'LQ'}}+\int_{L_2}+\int_{L_1}\right)\left(\frac{xf'(x)}{y[g(x)-\lambda]}\right)dy \\ &\qquad-\left(\int_{L_2}+\int_{L_1}\right)\left(\frac{xf'(x)}{y[g(x)-\lambda]}\right)dy \\ &\quad=\int\int_{D_1}\frac{1}{y}\frac{d}{dx}\left(\frac{xf'(x)}{g(x)-\lambda}\right)dx\,dy-\int_{y_2'}^{y_1'}\frac{x_1'f_1'(x_1')}{y[g(x_1')-\lambda]}\,dy \\ &\quad<-\int_{y_2}^{y_1}\frac{x_1'f_1'(x_1')}{y[g(x_1')-\lambda]}\,dy\quad\text{(by assumption (iv)).}\end{aligned} \tag{54}$$

Now we can consider the integration along the arc $\widehat{QRP}$. Let $L_1'=TL_1$, $L_2'=TL_2$ and let D_2 be the domain bounded by the arc $\widehat{QRP}$ of Γ, L_1', and L_2'. Then

$$\begin{aligned}\int_{\widehat{QRP}}xf'(x)\,dt&=\left(\int_{QRP}+\int_{-L_1'}+\int_{-L_2'}\right)\frac{xf'(x)}{y[g(x)-\lambda]}\,dy \\ &\qquad-\left(\int_{-L_1'}+\int_{-L_2'}\right)\frac{xf'(x)}{y[g(x)-\lambda]}\,dy \\ &=\int\int_{D_2}\frac{1}{y}\frac{d}{dx}\left(\frac{xf'(x)}{g(x)-\lambda}\right)dx\,dy+\int_{y_2}^{y_1}\frac{x_1f_2'(x_1)}{y[g(x_1)-\lambda]}\,dy \\ &<\int_{y_2}^{y_1}\frac{x_1f_2'(x_1)}{y[g(x_1)-\lambda]}\,dy.\end{aligned} \tag{55}$$

Combining (54) and (55), we have

$$\left(\int_{\widehat{P'LQ'}} + \int_{\widehat{QRP}}\right)(xf'(x))\,dt < \int_{y_2}^{y_1}\left[\frac{x_1 f_2'(x_1)}{g(x_1)-\lambda} - \frac{x_1' f_1'(x_1')}{g(x_1')-\lambda}\right]\frac{dy}{y} \tag{56}$$
$$< 0 \quad \text{(by (7))}.$$

Combining (49), (53), and (56), we have

$$\oint_\Gamma \operatorname{Div}(h, k)\,dt = \oint_\Gamma xf'(x)\,dt < 0. \tag{57}$$

This completes the proof of this lemma. □

Now we are in a position to prove Theorem 1.

Proof of Theorem 1. From Lemma 1, the solutions are positive and bounded. From Lemma 2, the equilibrium point (x^*, y^*) is a source. Hence there exists a closed orbit. But from Lemmas 3, 4, and 5, each closed orbit must be stable. But two adjacent periodic orbits cannot be positively stable on the sides facing each other (Coddington and Levinson [4, Thm. 3.4, p. 397]). Hence the closed orbit is a unique limit cycle. It is easy to see that this limit cycle is also globally stable, that is, nonequilibrium solutions will tend to this cycle eventually. This completes the proof of Theorem 1. □

Remark. In the proof of Lemma 5, we introduce the line segments L_1 and L_2. In the original proof of Cheng [2], we use the line segment $\overline{P'Q'}$ instead. Hadeler pointed out to us that $\overline{P'Q'}$ may intersect the orbit Γ [5]. This is the gap (in [2]) mentioned in the Introduction.

Acknowledgments. K.-S. Cheng thanks K. P. Hadeler for pointing out the gap referred to in the above remark. Both authors express thanks to Paul Waltman and Sze-Bi Hsu for their constant interest in this problem and their kind encouragement.

REFERENCES

[1] F. ALBRECHT, H. GATZKE, A. HADDAD, AND N. WAX, *The dynamics of two interacting populations*, J. Math. Anal. Appl., 46 (1974), pp. 658-670.

[2] K. S. CHENG, *Uniqueness of a limit cycle for a predator-prey system*, SIAM J. Math. Anal., 12 (1981), pp. 541-548.

[3] K. S. CHENG, S. B. HSU, AND S. S. LIN, *Some results on global stability of a predator-prey system*, J. Math. Biol., 12 (1981), pp. 115-126.

[4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[5] K. P. HADELER, Private communication.

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

[7] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1973.

[8] S. B. HSU, S. P. HUBBELL, AND P. WALTMAN, *Competing predators*, SIAM J. Appl. Math., 35 (1978), pp. 617-625.

[9] ———, *A contribution to the theory of competing predators*, Ecological Monographs, 48 (1978), pp. 337-349.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

009

A DIRECT LYAPUNOV APPROACH TO VOLTERRA INTEGRODIFFERENTIAL EQUATIONS*

OLOF J. STAFFANS†

Abstract. We propose a different approach to the Lyapunov theory for Volterra integrodifferential equations. Instead of using Lyapunov functionals we use Lyapunov functions, which typically are the norm of the solution raised to some power. Our Lyapunov functions are not decreasing along solutions, so we use separate estimates to bound them from above. Our approach is direct and straightforward, and it appears to be applicable to most of the results that have been proved earlier by means of Lyapunov functionals.

Key words. Lyapunov, Volterra equation

AMS(MOS) subject classifications. 45D05, 45A05, 45M10

1. Introduction. An important ingredient in the stability theory for ordinary differential equations (ODEs) is Lyapunov's method. Theoretically this method is very appealing, and there are applications where it is natural to use it.

For the convenience of the reader, let us give a short description of Lyapunov's method for the autonomous ordinary differential equation

$$x'(t)=f(x(t)), \quad t\geqq 0, \quad x(0)=x_0. \tag{1.1}$$

We denote the solution of (1.1) with initial condition $x(0)=\xi$ by $x(t,\xi)$ (assuming that a unique solution exists for each ξ), and define

$$\dot{V}(\xi)=\limsup_{h\to 0+}\frac{1}{h}[V(x(h,\xi))-V(\xi)].$$

Clearly, under sufficient differentiability assumptions, we have $\dot{V}(\xi)=\langle V'(\xi),f(\xi)\rangle$, where V' represents the gradient of V.

The following result is found in [28, Thm. 1.1, p. 293].

PROPOSITION 1.1. *If there is a positive definite function V on Ω with $\dot{V}\leqq 0$, then the solution $x=0$ of (1.1) is stable. If, in addition, $-\dot{V}$ is positive definite on Ω, then the solution $x=0$ is asymptotically stable.*

The function V above is called a *Lyapunov function* for (1.1) on Ω.

A companion result also exists, which says that if $\dot{V}$ is positive definite, then the zero solution is unstable (to prove this it suffices to observe that by Proposition 1.1, the equation is asymptotically stable in the backwards time direction). Actually, it is possible to prove instability under somewhat weaker assumptions; see [28, Thm. 1.2, p. 294].

Numerous attempts have veen made to obtain similar results for integral and functional equations. The formal theory presents no difficulties. However, it is a quite difficult task to find a Lyapunov function for a given ordinary differential equation, and it is virtually impossible to find a Lyapunov function or functional for a Volterra integral or a functional equation. There are some simple exceptions to this general rule, and we shall discuss these exceptions below.

The key requirement in Proposition 1.1 is that V is nonincreasing along solutions. In practice this requirement is so difficult to satisfy that to many ordinary differential

* Received by the editors February 5, 1987; accepted for publication June 4, 1987.
† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

equations we cannot apply Lyapunov theory unless we know in advance that the zero solution is asymptotically stable. Of course, if we know in advance that the zero solution is asymptotically stable, then there is no need to apply Lyapunov theory.

The situation becomes even worse when we replace the ordinary differential equation with an integral equation. If we want to have a Lyapunov function V which is nonincreasing along the solutions of the equation, then, due to the way in which solutions of integral equations behave, we are forced to let V depend not only on the present value $x(t)$ of x, but also on past values $x(s)$ with $s \leqq t$. In other words, we use Lyapunov functionals instead of Lyapunov functions. This makes it more difficult to specify what we mean by the second requirement in Proposition 1.1, i.e., the requirement that V should be positive definite.

During the last eight years, Burton, Huang, and Mahfoud have developed a Lyapunov theory, which primarily seems to apply to Volterra integrodifferential equations that have a dominant ODE part, or more generally, equations that can be transformed into equations with a dominant ODE part (see [2]–[17]). They use Lyapunov functionals, which are (most of the time) nonincreasing or strictly decreasing along solutions. The purpose of this work is to show that we can obtain the same results by using Lyapunov functions, which are allowed to increase as well as decrease, but stay bounded from above and below. Our approach applies to the same general class of equations, and it has two advantages:

(1) It is easy to construct the Lyapunov function. More specifically, we use the same Lyapunov function for the integrodifferential equation as we do for the corresponding unperturbed ordinary differential equation.

(2) The estimates that we need to compensate for the fact that our Lyapunov function is allowed to increase as well as decrease are simple, most of the time completely trivial (one uses the variation of constants formula, or a comparison theorem, or integration by parts).

The reader may object to the fact that most of the examples that we discuss below are fairly simple, and that they may be regarded as small perturbations of ordinary differential equations. This is true. However, apart from a number of well-known results for equations with kernels of positive type, these are the only ones that fall within the scope of this note, because they are the only ones that we know about to which the Lyapunov theory has been successfully applied. In addition, as several examples given in [16] and [17] show, this class of equations is larger than what we, at first sight, might expect.

The technique that we use is a variant of the so-called *energy technique.* This means that we obtain most of our basic estimates (i.e., those that we call L^1-estimates and L^2-estimates) by taking the inner product of the equation and some function, and integrating.

To keep this paper at a reasonable length, we restrict ourselves to linear Volterra integrodifferential equations, but the argument that we present can be applied to more general equations as well, e.g., to some nonlinear Volterra equations, and to nonlinear functional differential equations. For the same reason we, moreover, restrict ourselves to the case where we want to prove stability. However, it should not be too difficult for readers to convince themselves that the same type of argument can be applied when they want to prove instability.

2. The reduction to the scalar case. Below we show how one may reduce a system of equations to a scalar integrodifferential equation or inequality, or equivalently, how to construct Lyapunov functions for a system of integrodifferential equations. (The

Lyapunov function that we use is a natural Lyapunov function for the ordinary differential equation, which is the dominant part of the integrodifferential equation.)

Let us look at the equation

$$x'(t)+Ax(t)=\int_{t_0}^{t} C(t,s)x(s)\,ds+f(t), \quad t\geqq t_0, \quad x(t_0)=x_0. \tag{2.1}$$

Here x and f take their values in $\mathbf{R}^n$, A is an $n\times n$ matrix with real entries, $C(t,s)$ is a matrix-valued function, and t_0 is a real number.

The basic assumption in this equation is that the ordinary differential equation

$$x'(t)+Ax(t)=f(t), \quad t\geqq t_0, \quad x(t_0)=x_0, \tag{2.2}$$

is asymptotically stable. This means that it is possible to find a unique symmetric, positive definite matrix B such that

$$A^TB+BA=I$$

(see [10, Thm. 5.11, p. 124]). Let α be the smallest eigenvalue and β the largest eigenvalue of this matrix (an equivalent definition is $\beta=\|B\|$ and $\alpha=\|B^{-1}\|^{-1}$, where $\|\cdot\|$ is the matrix norm corresponding to the Euclidean norm in $\mathbf{R}^n$). Then $0<\alpha\leqq\beta$, and for all $x\in\mathbf{R}^n$,

$$\alpha\langle x,x\rangle\leqq\langle x,Bx\rangle\leqq\beta\langle x,x\rangle; \qquad \langle Bx,Bx\rangle\leqq\beta\langle x,Bx\rangle,$$

where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product.

In the sequel we denote the Euclidean norm in $\mathbf{R}^n$ by $|\cdot|$, and denote the corresponding matrix norm by $\|\cdot\|$. We let $|\cdot|_B$ be the norm induced by the inner product $\langle x,y\rangle_B=\langle x,By\rangle$, i.e., $|x|_B=|\sqrt{B}\,x|$. The corresponding matrix norm is denoted by $\|\cdot\|_B$. These norms satisfy

$$\sqrt{\alpha}\,|\cdot|\leqq|\cdot|_B\leqq\sqrt{\beta}\,|\cdot|, \qquad \sqrt{\alpha/\beta}\,\|\cdot\|\leqq\|\cdot\|_B\leqq\sqrt{\beta/\alpha}\,\|\cdot\|.$$

In our L^1 and L^∞ estimates we use the Lyapunov function

$$z(t)=|x(t)|_B, \qquad t\in[t_0,\infty).$$

This function is locally absolutely continuous, and its derivative $z'(t)$ satisfies (for almost all $t\geqq t_0$)

$$z'(t)=\frac{1}{|x(t)|_B}\left(-\frac{1}{2}|x(t)|^2+\int_{t_0}^{t}\langle x(t),BC(t,s)x(s)\rangle\,ds+\langle x(t),Bf(t)\rangle\right). \tag{2.3}$$

It is natural to do one of two things: Either replace the only remaining $|x(t)|$ by $|x(t)|_B$, or replace $|x(t)|_B$ by $|x(t)|$. The first alternative leads to the inequality

$$z'(t)+\frac{1}{2\beta}z(t)\leqq\int_{t_0}^{t}\|C(t,s)\|_Bz(s)\,ds+|f(t)|_B, \qquad t\geqq t_0, \tag{2.4}$$

and the second alternative leads to the inequality

$$z'(t)+\frac{1}{2\sqrt{\beta}}|x(t)|\leqq\int_{t_0}^{t}\|\sqrt{B}\,C(t,s)\|\,|x(s)|\,ds+|f(t)|_B, \qquad t\geqq t_0. \tag{2.5}$$

The first of these two equations has the advantage that it contains only one function $z(t)$ as opposed to the two functions $z(t)$ and $|x(t)|$ in (2.5), but as we shall see below, also the latter equation can be exploited. Note, in particular, that the smallest eigenvalue α of B does not enter in the expressions above (unless we estimate $\|C(t,s)\|_B$ by $\sqrt{\beta/\alpha}\|C(t,s)\|$).

Equation (2.1) is a perturbation of a time-independent ordinary differential equation. In a similar way we can perturb a time-dependent differential equation to get

$$x'(t)+A(t)x(t)=\int_{t_0}^{t} C(t,s)x(s)\,ds+f(t),\quad t\geqq t_0,\quad x(t_0)=x_0. \tag{2.6}$$

The setting is the same as before, except that we let $A\in L^1_{\mathrm{loc}}([t_0,\infty);\mathbf{R}^{n\times n})$. We suppose that it is possible to find a positive definite matrix B such that the matrix

$$R(t)=A^T(t)B+BA(t)$$

is positive definite for almost all $t\geqq t_0$. Define α and β as before, and define

$$p(t)=\beta\inf_{x\in\mathbf{R}^n}\frac{\langle x,R(t)x\rangle}{|x|_B^2},\qquad q(t)=\sqrt{\beta}\inf_{x\in\mathbf{R}^n}\frac{\langle x,R(t)x\rangle}{|x|_B|x|}$$

(here β and $\sqrt{\beta}$ are scaling factors which make $p(t)=q(t)=1$ if $R(t)=I$). (The constant $p(t)$ could also have been defined as the smallest eigenvalue of the matrix $\beta\sqrt{B}\,R(t)(\sqrt{B})^{-1}$.) It is not difficult to check that

$$p(t)\geqq q(t)\geqq\sqrt{\alpha/\beta}\,p(t).$$

In this case (2.3) becomes

$$z'(t)=\frac{1}{|x(t)|_B}\left(-\frac{1}{2}\langle x(t),R(t)x(t)\rangle+\int_{t_0}^{t}\langle x(t)C(t,s),Bx(s)\rangle\,ds+\langle x(t),Bf(t)\rangle\right), \tag{2.7}$$

and the inequalities (2.4) and (2.5) become

$$z'(t)+\frac{p(t)}{2\beta}z(t)\leqq\int_{t_0}^{t}\|C(t,s)\|_B z(s)\,ds+|f(t)|_B,\qquad t\geqq t_0, \tag{2.8}$$

and

$$z'(t)+\frac{q(t)}{2\sqrt{\beta}}|x(t)|\leqq\int_{t_0}^{t}\|\sqrt{B}\,C(t,s)\|\,|x(s)|\,ds+|f(t)|_B,\qquad t\geqq t_0. \tag{2.9}$$

Observe that there is nothing that prevents us from choosing $B=I$. Likewise, there is nothing that prevents us from applying this approach to the time-independent case. If we in the time-independent case choose $B=I$, then $R=A^T+A$, $\alpha=\beta=1$, and $p(t)=q(t)$ is the smallest eigenvalue of A^T+A (which is required to be positive, if we are to choose B this way).

It is possible to use a time-dependent transformation matrix $B(t)$ as well. In this case the Lyapunov function z is

$$z(t)=\langle x(t),B(t)x(t)\rangle^{1/2},$$

and we redefine R to be

$$R(t)=A^T(t)B(t)+B(t)A(t)-B'(t).$$

We leave the details to the reader (cf. [35, §§ 6.5, 6.6, pp. 114–127]).

In the scalar case we throughout choose $B=I$, $\alpha=\beta=1$ and $p(t)=q(t)=2A(t)$.

3. Some L^1-estimates. Let us first develop the most elementary estimate for (2.8) and (2.9) (the same estimate is of course valid for (2.4) and (2.5) if one replaces $p(t)$

and $q(t)$ by one). To get this estimate we integrate (2.8) and (2.9) over $[t_0, T]$, and change the order of integration. This leads to the inequalities

$$z(T)+\int_{t_0}^{T}\left(\frac{p(s)}{2\beta}-\int_s^T \|C(t,s)\|_B\,dt\right)z(s)\,ds \leqq z(t_0)+\int_{t_0}^{T}|f(t)|_B\,dt, \qquad T\geqq t_0, \tag{3.1}$$

and

$$\begin{aligned} &z(T)+\int_{t_0}^{T}\left(\frac{q(s)}{2\sqrt{\beta}}-\int_s^T \|\sqrt{B}\,C(t,s)\|\,dt\right)|x(s)|\,ds \\ &\qquad \leqq z(t_0)+\int_{t_0}^{T}|f(t)|_B\,dt, \qquad T\geqq t_0. \end{aligned} \tag{3.2}$$

From these inequalities we get by direct inspection the first part of the following theorem.

THEOREM 3.1. (i) *If $f\in L^1([t_0,\infty);\mathbf{R}^n)$, and $\int_s^\infty \|C(t,s)\|_B\,dt\leqq p(s)/2\beta$ for almost all $s\geqq t_0$, or $\int_s^\infty \|\sqrt{B}\,C(t,s)\|\,dt\leqq q(s)/2\sqrt{\beta}$ for almost all $s\geqq t_0$, then the solution x of* (2.6) *is bounded. More precisely, $|x(t)|_B\leqq |x_0|_B+\int_{t_0}^t |f(s)|_B\,ds$ for all $t\geqq t_0$.*

(ii) *Let $f\in L^1([t_0,\infty);\mathbf{R}^n)$, and assume that for some constant $\varepsilon\in(0,1/2\beta)$, $\int_s^\infty \|C(t,s)\|_B\,dt\leqq (p(s)/2\beta)-\varepsilon(1+p(s))$ for almost all $s\geqq t_0$, or $\int_s^\infty \|\sqrt{B}\,C(t,s)\|\,dt\leqq (q(s)/2\sqrt{\beta})-\varepsilon(1+p(s))$ for almost all $s\geqq t_0$. Then the function $z(\cdot)=|x(\cdot)|_B$ satisfies $\int_{t_0}^\infty ((1+p(t))z(t)+|z'(t)|)\,dt<\infty$. In particular, $x(t)\to 0$ as $t\to\infty$.*

(iii) *In addition to* (ii), *suppose that $\|A(s)\|\leqq M(1+p(s))$ for some constant M and almost all $s\geqq t_0$. Then $x'\in L^1([t_0,\infty);\mathbf{R}^n)$.*

Note, in particular, that in the case where A is unbounded we do not have to assume that $\operatorname{ess\,sup}_{s\geqq t_0}\int_s^\infty \|C(t,s)\|\,dt<\infty$. In this respect Theorem 3.1 and some of the other theorems below seem to be new.

The second and third conclusions can be simplified slightly in the autonomous case (2.1). Then the additional assumption in (iii) is automatically satisfied, and the conditions $\int_s^\infty \|C(t,s)\|_B\,dt\leqq (p(s)/2\beta)-\varepsilon(1+p(s))$ and $\int_s^\infty \|\sqrt{B}\,C(t,s)\|\,dt\leqq (q(s)/2\sqrt{\beta})-\varepsilon(1+p(s))$ can be written as $\int_s^\infty \|C(t,s)\|_B\,dt\leqq (1/2\beta)-\varepsilon$ and $\int_s^\infty \|\sqrt{B}\,C(t,s)\|\,dt\leqq (1/2\sqrt{\beta})-\varepsilon$.

The additional assumption in (iii) is automatically satisfied in the scalar time-dependent case as well.

Proof. As we already observed above, the first of the two claims follows from a direct inspection of (3.1) and (3.2). Likewise, it is a direct consequence of (3.1) and (3.2) that under the extra assumption we have $\int_{t_0}^\infty (1+p(t))z(t)\,dt<\infty$. This implies that the last two terms on the right-hand side of (2.7) are integrable (estimate these terms in the same say as in (2.8)). Since z is bounded (from below), and the first term on the right-hand side of (2.7) is nonpositive, also this term must be integrable. This proves that z' is integrable. That x' itself is integrable under the additional assumption made in (iii) follows directly from (2.6). □

Some scalar nonlinear convolution versions of this result are given in [33, Thm. 1, p. 458] and [34, Thm. 1, p. 340], and some nonlinear nonconvolution versions are given in [19]–[21].

Theorem 3.1 contains and extends the claims about stability and asymptotic stability made in [2, Thm. 1, p. 102], [2, Thm. 2, p. 104], [3, Thm. 1, p. 42], and [3, Thm. 2, p. 43]. (We shall return below to the question of uniform stability and uniform asymptotic stability.)

There is another equally obvious estimate which can be applied to (2.8) and (2.9). Instead of just integrating (2.8) we introduce a scalar function η (which need not be

positive), and add and subtract the term $\eta(t)z(t)$ from (2.8). Then we multiply the equation by $\exp\left(\int_{t_0}^{t}\eta(v)\,dv\right)$, integrate over $[t_0, T]$, and finally divide by $\exp\left(\int_{t_0}^{T}\eta(v)\,dv\right)$. This leads to the inequality

$$z(T)+\int_{t_0}^{T}\left(\frac{p(s)}{2\beta}-\eta(s)-\int_{s}^{T}\exp\left(\int_{s}^{t}\eta(v)\,dv\right)\|C(t,s)\|_B\,dt\right)$$

$$\times\exp\left(-\int_{s}^{T}\eta(v)\,dv\right)z(s)\,ds \tag{3.3}$$

$$\leqq z(t_0)\exp\left(-\int_{t_0}^{T}\eta(v)\,dv\right)+\int_{t_0}^{T}\exp\left(-\int_{t}^{T}\eta(v)\,dv\right)|f(t)|_B\,dt, \qquad T\geqq t_0.$$

The same manipulations applied to (2.9), with $\eta\geqq 0$, lead to the inequality

$$z(T)+\int_{t_0}^{T}\left(\frac{q(s)}{2\sqrt{\beta}}-\sqrt{\beta}\,\eta(s)-\int_{s}^{T}\exp\left(\int_{s}^{t}\eta(v)\,dv\right)\|\sqrt{B}\,C(t,s)\|\,dt\right)$$

$$\times\exp\left(-\int_{s}^{T}\eta(v)\,dv\right)|x(s)|\,ds \tag{3.4}$$

$$\leqq z(t_0)\exp\left(-\int_{t_0}^{T}\eta(v)\,dv\right)+\int_{t_0}^{T}\exp\left(-\int_{t}^{T}\eta(v)\,dv\right)|f(t)|_B\,dt, \qquad T\geqq t_0.$$

This very last estimate gives us the following theorem.

THEOREM 3.2. *If*

$$\int_{s}^{\infty}\exp\left(\int_{s}^{t}\eta(v)\,dv\right)\|\sqrt{B}\,C(t,s)\|\,dt\leqq\frac{q(s)}{2\sqrt{\beta}}-\sqrt{\beta}\,\eta(s)$$

for some function $\eta\geqq 0$ and almost all $s\geqq t_0$, then the solution x of (2.6) *satisfies*

$$|x(t)|_B\leqq|x_0|_B\exp\left(-\int_{t_0}^{t}\eta(v)\,dv\right)+\int_{t_0}^{t}\exp\left(-\int_{s}^{t}\eta(v)\,dv\right)|f(s)|_B\,ds$$

for all $t\geqq t_0$. In particular, if the right-hand side of this inequality is bounded, then so is x, and if it tends to zero as $t\to\infty$, then so does x.

The proof is obvious.

A scalar nonlinear version of this result is given in [20, Cor. 2.4, pp. 328-329].

Theorem 3.2 extends [2, Thm. 3, p. 106], [3, Thm. 4, p. 45], [6, Cor. 1.3, p. 173], [6, Thm. 3, p. 180], and [8, Thm. 3, p. 278]. In these theorems Burton uses various conditions, some of which do not directly seem to be related to the assumption of Theorem 3.2. However, they are all special cases of Theorem 3.2. For example, in formula (2.5.21) of [10, p. 44], it is assumed that

$$\|C(t,s)\|\geqq\lambda(t)\int_{t}^{\infty}\|C(u,s)\|\,du, \tag{3.5}$$

where λ is nonnegative. If we choose $\eta(t)=\varepsilon\lambda(t)$ for some small positive ε, then it is easy to show (integrate by parts) that

$$\int_{s}^{\infty}\exp\left(\int_{s}^{t}\eta(v)\,dv\right)\|C(t,s)\|\,dt\leqq\frac{1}{1-\varepsilon}\int_{s}^{\infty}\|C(t,s)\|\,dt.$$

Clearly, this means that if, in addition to (3.5), one has $\int_{s}^{\infty}\|C(t,s)\|dt\leqq(1-\varepsilon)(q(s)/2\beta-\varepsilon\lambda(s))$ for almost all $s\geqq t_0$, then Theorem 3.2 applies (recall that $\|\sqrt{B}\,C(t,s)\|\leqq\|\sqrt{B}\|\,\|C(t,s)\|\leqq\sqrt{B}\|C(t,s)\|$).

A result similar to Theorem 3.2 is true for the inequality (3.3) as well, but we leave the formulation of that result to the reader.

Another possible use of (3.3) is the following: Let us introduce the notation

$$|z|_- = \max[0, -z], \qquad z \in \mathbf{R},$$

and, let us define λ by

$$\lambda \overset{\text{def}}{=} \sup_{T \geqq t_0} \int_{t_0}^{T} \left| \frac{p(s)}{2\beta} - \eta(s) - \int_s^T \exp\left(\int_s^t \eta(v)\, dv\right) \|C(t,s)\|_B\, dt \right|_- \times \exp\left(-\int_s^T \eta(v)\, dv\right) ds. \tag{3.6}$$

If we suppose that $z(T) = \max_{s \in [t_0, T]} z(s)$, then clearly (3.3) implies that

$$(1-\lambda) z(T) \leqq z(t_0) \exp\left(-\int_{t_0}^T \eta(v)\, dv\right) + \int_{t_0}^T \exp\left(-\int_t^T \eta(v)\, dv\right) |f(t)|_B\, dt.$$

This proves the following theorem.

THEOREM 3.3. *Suppose that the constant λ defined in* (3.6) *satisfies $\lambda < 1$, that*

$$\inf_{t \geqq t_0} \int_{t_0}^t \eta(v)\, dv > -\infty,$$

and that

$$\sup_{t \geqq t_0} \int_{t_0}^t \exp\left(-\int_s^t \eta(v)\, dv\right) |f(s)|_B\, ds < \infty.$$

Then the solution x of (2.6) *is bounded.*

This theorem extends [11, Prop., p. 247].

One particular case where Theorem 3.3 can be applied is the following. We integrate by parts to get

$$\int_s^T \exp\left(\int_s^t \eta(v)\, dv\right) \|C(t,s)\|_B\, dt = \int_s^T \|C(t,s)\|_B\, dt + \int_s^T \eta(t) \exp\left(\int_s^t \eta(v)\, dv\right) \int_t^T \|C(u,s)\|_B\, du\, dt.$$

Thus, if we assume that $\int_s^\infty \|C(t,s)\|_B\, dt \leqq (p(s)/2\beta) - \eta(s)$, and that $\eta \geqq 0$, then we can estimate the constant λ in (3.6) by

$$\begin{aligned}
\lambda &\leqq \int_{t_0}^T \int_s^T \eta(t) \exp\left(-\int_t^T \eta(v)\, dv\right) \int_t^T \|C(u,s)\|_B\, du\, dt\, ds \\
&= \int_{t_0}^T \eta(t) \exp\left(-\int_t^T \eta(v)\, dv\right) \int_{t_0}^t \int_t^T \|C(u,s)\|_B\, du\, ds\, dt \\
&\leqq \int_{t_0}^T \eta(t) \exp\left(-\int_t^T \eta(v)\, dv\right) dt \sup_{t \geqq t_0} \int_{t_0}^t \int_t^\infty \|C(u,s)\|_B\, du\, ds \\
&\leqq \sup_{t \geqq t_0} \int_{t_0}^t \int_t^\infty \|C(u,s)\|_B\, du\, ds.
\end{aligned}$$

Thus, if we define ρ by

$$\rho = \sup_{t \geqq t_0} \int_{t_0}^t \int_t^\infty \|C(u,s)\|_B\, du\, ds, \tag{3.7}$$

then $\lambda \leqq \rho$.

The argument above, combined with Theorem 3.3, proves the following theorem.

THEOREM 3.4. *Suppose that* $\int_s^\infty \|C(t,s)\|_B\,dt \leqq p(s)/2\beta - \eta(s)$ *for some function* $\eta \geqq 0$ *and almost all* $s \geqq t_0$. *In addition, suppose that the constant* ρ *defined in* (3.7) *satisfies* $\rho < 1$, *and that* $\sup_{t \geqq t_0} \int_{t_0}^t \exp\left(-\int_s^t \eta(v)\,dv\right)|f(s)|_B\,ds < \infty$. *Then the solution* x *of* (2.6) *is bounded.*

Theorem 3.4 extends [3, Thm. 3, p. 44], [6, Cor. 1.1, p. 171], [6, Cor. 1.2, p. 172], [7, Prop. 2, p. 64], and [7, Prop. 3, p. 66]. (In these results, $\eta(t)$ is a small constant.)

4. Some L^∞-estimates. Our L^∞-estimates are based on the following lemma.

LEMMA 4.1. *Let* z *be real-valued and locally absolutely continuous, and define the set* E *by* $E = \{t \in [t_0, \infty) \mid z(t) = \max_{s \in [t_0, t]}\}$. *Then the function* $t \mapsto \sup_{s \in [t_0,t]} z(s)$ *is locally absolutely continuous, and its derivative is almost everywhere equal to* $\chi_E(t)z'(t)$. *In particular, if* $z'(t) \leqq f(t)$ *for almost all* $t \in E$, *then* $z(t) \leqq z(t_0) + \int_{t_0}^t \chi_E(s)f(s)\,ds$ *for all* $t \geqq t_0$.

The easy proof of this lemma is left to the reader.

When we apply Lemma 4.1 to the inequality (2.8) we get the following result.

THEOREM 4.2. (i) *If* $f \in L^1([t_0, \infty); \mathbf{R}^n)$, *and* $\int_{t_0}^t \|C(t,s)\|_B\,ds \leqq p(t)/2\beta$ *for almost all* $t \geqq t_0$, *then the solution* x *of* (2.6) *is bounded. More precisely,* $|x(t)|_B \leqq |x_0|_B + \int_{t_0}^t |f(s)|_B\,ds$ *for all* $t \geqq t_0$.

(ii) *If* $\operatorname{ess\,sup}_{s \geqq t_0} (1+p(s))^{-1}|f(s)|_B < \infty$, *and* $\int_{t_0}^t \|C(t,s)\|_B\,ds \leqq p(t)/2\beta - \varepsilon(1+p(t))$ *for some constant* $\varepsilon \in (0, 1/2\beta)$ *and almost all* $t \geqq t_0$, *then the function* $z(\cdot) = |x(\cdot)|_B$ *satisfies* $z(t) \leqq \max\{|x_0|_B, (1/\varepsilon) \operatorname{ess\,sup}_{s \in [t_0,t]} (1+p(s))^{-1}|f(s)|_B\}$ *for all* $t \geqq t_0$, *and* $\operatorname{ess\,sup}_{t \geqq t_0} (1+p(t))^{-1}|z'(t)| < \infty$.

(iii) *In addition to* (ii), *suppose that* $\|A(s)\| \leqq M(1+p(s))$ *for some constant* M *and almost all* $s \geqq t_0$. *Then* $\operatorname{ess\,sup}_{t \geqq t_0} (1+p(t))^{-1}|x'(t)| < \infty$.

The proof is left to the reader (note that in (ii) the derivative will be nonpositive on E as soon as z has exceeded the given bound).

Theorem 4.2 generalizes [22, Thm. 4, p. 149].

By applying Lemma 4.1 to the function $\exp\left(\int_{t_0}^t \eta(v)\,dv\right)z(t)$ rather than to z itself we get the following analogue of Theorem 3.2.

THEOREM 4.3. *If*

$$\int_{t_0}^t \exp\left(\int_s^t \eta(v)\,dv\right)\|C(t,s)\|_B\,ds \leqq \frac{p(t)}{2\beta} - \eta(t)$$

for almost all $t \geqq t_0$, *then the solution* x *of* (2.6) *satisfies*

$$|x(t)|_B \leqq |x_0|_B \exp\left(-\int_{t_0}^t \eta(v)\,dv\right) + \int_s^t \exp\left(-\int_s^t \eta(v)\,dv\right)|f(s)|_B\,ds$$

for all $t \geqq t_0$. *In particular, if the right-hand side of this inequality is bounded, then so is* x, *and if it tends to zero as* $t \to \infty$, *then so does* x.

The easy proof is left to the reader.

Theorem 4.3 is closely related to [27, Thm. 3.1, p. 1398].

Extensive work has been done on the question of whether the function x in Theorem 4.2 tends to a limit at infinity. See [1, Thm. 4.2, p. 242], [25, Thm. 5.3, p. 264], [26, Cor. 1, p. 100], and [26, Thm. 3.1, p. 108].

5. L^2-estimates and kernels of positive type. To get an L^2-estimate rather than an L^1-estimate or an L^∞-estimate we use the Lyapunov function $w(t) = \frac{1}{2}z^2(t)$ instead of the function $z(t)$. From (2.7) we get

$$(5.1)\quad z(t)z'(t) = -\frac{1}{2}\langle x(t), R(t)x(t)\rangle + \int_{t_0}^t \langle x(t), BC(t,s)x(s)\rangle\,ds + \langle x(t), B(f(t)\rangle,$$

and (3.1) and (3.2) are replaced by

$$\begin{aligned}(5.2)\qquad &\frac{1}{2}z^2(T)+\int_{t_0}^{T}\frac{p(t)}{2\beta}z^2(t)\,dt-\int_{t_0}^{T}\int_{t_0}^{t}z(t)\|C(t,s)\|_B z(s)\,ds\,dt\\ &\leqq\frac{1}{2}z^2(t_0)+\int_{t_0}^{T}z(t)|f(t)|_B\,dt,\qquad T\geqq t_0,\end{aligned}$$

and

$$\begin{aligned}(5.3)\qquad &\frac{1}{2}z^2(T)+\int_{t_0}^{T}\frac{r(t)}{2}|x(t)|^2\,dt-\int_{t_0}^{T}\int_{t_0}^{t}|x(t)|\,\|BC(t,s)\|\,|x(s)|ds\,dt\\ &\leqq\frac{1}{2}z^2(t_0)+\int_{t_0}^{T}z(t)|f(t)|_B\,dt,\qquad T\geqq t_0,\end{aligned}$$

where

$$r(t)=\inf_{x\in\mathbf{R}^n}\frac{\langle x,R(t)x\rangle}{|x|^2}$$

is the smallest eigenvalue of the matrix $R(t)$. The double integral in (5.2) can be estimated in the following way: As the geometric mean is no larger than the arithmetic mean, we get for all $\kappa>0$,

$$\begin{aligned}(5.4)\qquad \int_{t_0}^{T}\int_{t_0}^{t}z(t)\|C(t,s)\|_B z(s)\,ds\,dt&\leqq\int_{t_0}^{T}\int_{t_0}^{t}\|C(t,s)\|_B\left(\frac{\kappa}{2}z^2(t)+\frac{1}{2\kappa}z^2(s)\right)ds\,dt\\ &=\int_{t_0}^{T}\left(\frac{\kappa}{2}\int_{t_0}^{t}\|C(t,u)\|_B\,du\right.\\ &\qquad\left.+\frac{1}{2\kappa}\int_{t}^{T}\|C(v,t)\|_B\,dv\right)z^2(t)\,dt.\end{aligned}$$

This is the key estimate in the remainder of this section.

Applying (5.4) to (5.2) and (5.3) we get the following result.

THEOREM 5.1. (i) *If $f\in L^1([t_0,\infty);\mathbf{R}^n)$, and there is some $\kappa>0$ such that*

$$\frac{\kappa}{2}\int_{t_0}^{t}\|C(t,u)\|_B\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|C(v,t)\|_B\,dv\leqq\frac{p(t)}{2\beta}$$

for almost $t>t_0$, or

$$\frac{\kappa}{2}\int_{t_0}^{t}\|BC(t,u)\|\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|BC(v,t)\|\,dv\leqq\frac{r(t)}{2}$$

for almost all $t>t_0$, then the solution x of (2.6) is bounded. More precisely, $|x(t)|_B\leqq|x_0|_B+\int_{t_0}^{t}|f(s)|_B\,ds$ for all $t\geqq t_0$.

(ii) *If $\int_{t_0}^{\infty}(1+p(t))^{-1}|f(t)|_B^2\,dt<\infty$, and there are constants $\kappa>0$ and $\varepsilon\in(0,1/2\beta)$ such that*

$$\frac{\kappa}{2}\int_{t_0}^{t}\|C(t,u)\|_B\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|C(v,t)\|_B\,dv\leqq\frac{p(t)}{2\beta}-\varepsilon(1+p(t))$$

for almost all $t>t_0$, or

$$\frac{\kappa}{2}\int_{t_0}^{t}\|BC(t,u)\|\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|BC(v,t)\|\,dv\leqq\frac{r(t)}{2}-\varepsilon(1+p(t))$$

for almost all $t>t_0$, then the function $w(\cdot)=|x(\cdot)|_B^2$ satisfies $\int_{t_0}^{\infty}((1+p(t))w(t)+|w'(t)|)\,dt<\infty$. In particular, $x(t)\to0$ as $t\to\infty$.

(iii) *In addition to* (ii), *suppose that* $\|A(s)\| \leqq M(1+p(s))$ *for some constant* M *and almost all* $s \geqq t_0$. *Then* $\int_{t_0}^{\infty} (1+p(t))^{-1}|x'(t)|^2\,dt < \infty$.

Proof. It follows from (5.2), (5.3), and (5.4) (and from the analogue of (5.4) for (5.3)) that in part (i) of the theorem,

$$\frac{1}{2}z^2(T) \leqq \frac{1}{2}z^2(t_0) + \int_{t_0}^{T} z(t)|f(t)|_B\,dt, \qquad T \geqq t_0. \tag{5.5}$$

Using, e.g., [36, Thm. 6.1, p. 121] we find that z is dominated by the solution y of the equation

$$\frac{1}{2}y^2(T) = \frac{1}{2}z^2(t_0) + \int_{t_0}^{T} y(t)|f(t)|_B\,dt, \qquad T \geqq t_0.$$

However, this is an ordinary differential equation, which can be solved explicitly to give the estimate in part (i).

In part (ii) of the theorem, instead of (5.5), we get from the same formulas as before (if necessary, decrease the value of ε, and replace $|x(t)|$ by $z(t)$)

$$\frac{1}{2}z^2(T) + \varepsilon\int_{t_0}^{T} (1+p(t))z^2(t)\,dt \leqq \frac{1}{2}z^2(t_0) + \int_{t_0}^{T} z(t)|f(t)|_B\,dt, \qquad T \geqq t_0.$$

This together with Hölder's inequality gives the conclusion that $\int_{t_0}^{\infty}(1+p(t))z^2(t)\,dt < \infty$. To prove the claim about the derivative one argues in the same way as in the proof of Theorem 3.1.

To prove (iii) one uses (2.6), observes that for some constant $M > 0$,

$$\operatorname*{ess\,sup}_{t \geqq t_0} (1+p(t))^{-1}\int_{t_0}^{t} \|C(t,s)\|_B\,ds \leqq M,$$

$$\operatorname*{ess\,sup}_{s \geqq t_0} (1+p(s))^{-1}\int_{s}^{\infty} \|C(t,s)\|_B\,dt \leqq M,$$

and uses Hölder's inequality and a change of the order of integration to get

$$\begin{aligned}
&\int_{t_0}^{T} (1+p(t))^{-1}\left[\int_{t_0}^{t} \|C(t,s)\|_B z(s)\,ds\right]^2 dt \\
&\quad \leqq \int_{t_0}^{T} (1+p(t))^{-1}\int_{t_0}^{t} \|C(t,s)\|_B\,ds \int_{t_0}^{t} \|C(t,s)\|_B z^2(s)\,ds\,dt \\
&\quad \leqq M\int_{t_0}^{T}\int_{s}^{T} \|C(t,s)\|_B\,dt\,z^2(s)\,ds \\
&\quad \leqq M^2\int_{t_0}^{T} (1+p(s))z^2(s)\,ds.
\end{aligned}$$

□

Theorem 5.1 contains the linear versions of the stability claims in [5, Thm. 1, p. 92], [5, Thm. 5, p. 100], [16, Thm. 3, p. 148], [16, Cor. 1, p. 149], and [16, Thm. 8, p. 156]. (The nonlinear versions given in [5] can be proved in a similar way.)

It is customary to make the following definition.

DEFINITION 5.2. A linear mapping F which takes the $\mathbf{R}^n$-valued function x into the $\mathbf{R}^n$-valued function $F(x)(t)$ is of positive type with respect to the inner product $\langle\cdot,\cdot\rangle_B$ on the interval $[t_0, \infty)$ if it is true for all continuous x and all $T > t_0$ that

$$\int_{t_0}^{T} \langle x(t), F(x)(t)\rangle_B \geqq 0.$$

Clearly, Theorem 5.1 is a special case of the following result (the proof remains the same).

THEOREM 5.3. (i) *If* $f \in L^1([t_0, \infty); \mathbf{R}^n)$, *and if the mapping which takes* x *into the function* $A(t)x(t) - \int_{t_0}^t C(t, s)x(s)\,ds$ *is of positive type with respect to the inner product* $\langle \cdot, \cdot \rangle_B$ *on* $[t_0, \infty)$, *then the solution* x *of* (2.6) *is bounded. More precisely,* $|x(t)|_B \leqq |x_0|_B + \int_{t_0}^t |f(s)|_B\,ds$ *for all* $t \geqq t_0$.

(ii) *If* $\int_{t_0}^\infty (1+p(t))^{-1}|f(t)|_B^2\,dt < \infty$, *and there is a constant* $\varepsilon \in (0, 1/2\beta)$ *such that the mapping that takes* x *into the function* $A(t)x(t) - \int_{t_0}^t C(t, s)x(s)\,ds - \varepsilon(1+p(t))x(t)$ *is of positive type with respect to the inner product* $\langle \cdot, \cdot \rangle_B$ *on* $[t_0, \infty)$, *then the solution* x *of* (2.6) *is bounded and satisfies* $\int_{t_0}^\infty (1+p(t))|x(t)|_B^2\,dt < \infty$. *If, in addition,* $\int_{t_0}^t \|C(t, u)\|_B\,du + \int_t^\infty \|C(v, t)\|_B\,dv \leqq M(1+p(t))$ *for some constant* $M > 0$ *and almost all* $t \geqq t_0$, *then the derivative of the function* $\|x(t)\|_B^2$ *is integrable. In particular, in this case* $x(t) \to 0$ *as* $t \to \infty$.

(iii) *In addition to* (ii), *suppose that* $\|A(s)\| \leqq M(1+p(s))$ *for some constant* M *and almost all* $s \geqq t_0$. *Then* $\int_{t_0}^\infty (1+p(t))^{-1}|x'(t)|^2\,dt < \infty$.

Scalar convolution versions of this result are given in [41, p. 83] and [43, Cor. 3.2, p. 132]. Some other results which fall within the scope of this theorem are [4, Thm. 3, p. 397], [5, Thm. 3, p. 97], and [5, Thm. 7, p. 103].

(Two of the theorems in [5], i.e., Theorem 2 on p. 94 and Theorem 6 on p. 102, can be proved with standard energy technique for second-order ordinary differential equations. It is easy to show that, under the assumptions in [5], the equations discussed in these two theorems are integrated versions of the damped nonlinear oscillator equation

$$x''(t) - \lambda(t)x'(t) + C(t, t)E(x(t)) = 0,$$

with initial condition $x'(0) = 0$.)

For additional results on scalar nonconvolution kernels of positive type, see [23] and the references mentioned there.

By adding and subtracting a term $\eta(t)z^2(t)$ from (5.1), multiplying the equation by $\exp(2\int_{t_0}^t \eta(v)\,dv)$, and finally integrating, we can prove the following analogue of Theorems 3.2 and 4.3.

THEOREM 5.4. *If the mapping that takes* x *into the function*

$$A(t)x(t) - \int_{t_0}^t \exp\left(\int_s^t \eta(v)\,dv\right) C(t, s)x(s)\,ds$$

is of positive type with respect to the inner product $\langle \cdot, \cdot \rangle_B$ *on* $[t_0, \infty)$, *then the solution* x *of* (2.6) *satisfies*

$$|x(t)|_B \leqq |x_0|_B \exp\left(-\int_{t_0}^t \eta(v)\,dv\right) + \int_{t_0}^t \exp\left(-\int_s^t \eta(v)\,dv\right)|f(s)|_B\,ds$$

for all $t \geqq t_0$. *In particular, if the right-hand side of this inequality is bounded, then so is* x, *and if it tends to zero as* $t \to \infty$, *then so does* x.

This result seems to be new.

Note, in particular, that if the size hypothesis on C in part (i) of Theorem 5.1 holds with $C(t, s)$ replaced by $\exp(\int_s^t \eta(v)\,dv)C(t, s)$, then Theorem 5.4 applies.

6. Perturbations of the derivative. So far, in our study of perturbed versions of the ordinary differential equation

$$x'(t) + A(t)x(t) = f(t), \qquad t \geqq t_0, \tag{6.1}$$

we have only permitted perturbations that can be dominated by the term $A(t)x(t)$ (Theorems 3.3 and 3.4 were exceptions; there we needed the term $x'(t)$ as well). Next we want to allow perturbations that are dominated by the term $x'(t)$. More precisely,

we study equations of the type

$$\frac{d}{dt}\left(x(t)+\int_{t_0}^{t} G(t,s)x(s)\,ds+g(t)\right)+A(t)x(t)=\int_{t_0}^{t} C(t,s)x(s)\,ds+f(t),$$
$$t\geqq t_0,\quad x(t_0)=x_0. \tag{6.2}$$

If G and g are sufficiently differentiable, then one can carry out the differentiations to get an equation which is of the form (2.6), with $A(t)$ replaced by $A(t)+G(t,t)$, $C(t,s)$ replaced by $C(t,s)-(\partial/\partial t)G(t,s)$ and $f(t)$ replaced by $f(t)-g'(t)$. In particular, if $A(t)+G(t,t)=0$, then the equation which one gets in this way has no leading ODE term. Therefore, we should not regard (6.2) as a perturbation of the equation

$$x'(t)+(A(t)+G(t,t))x(t)=f(t),\qquad t\geqq t_0. \tag{6.3}$$

Instead, it should be regarded as a perturbation of (6.1).

References [16] and [17] contain several nice examples on equations of the type (2.6), which can be rewritten in the form (6.2) in such a way that (6.2) becomes a small perturbation of (6.1).

Equations of the type (6.2) are a special subclass of the so-called neutral functional differential equations, and the operator which maps x into the function $t\mapsto x(t)+\int_{t_0}^{t} G(t,s)x(s)\,ds+g(t)$ is usually called the D-operator. Under the assumptions which we use below, both the D-operator and the unperturbed equation (6.1) will be asymptotically stable. (For a further discussion on a neutral functional differential equation with a stable D-operator, see [44].)

There are at least two different ways to approach (6.2). One possibility is to first use the variation of constants formula for (6.1), and then apply the contraction mapping principle. We shall return to this approach in the next section. The other possibility, the one we present now, is to use L^2-estimates in the spirit of the preceding section.

Earlier, when we developed our L^2-estimates for (2.6), we simply took the B-inner product of the equation with $x(t)$, and integrated. This time, the natural thing to do is to take the B-inner product of (6.2) with $x(t)+\int_{t_0}^{t} G(t,s)x(s)\,ds+g(t)$, and integrate. This leads to the identity (we have collected all the quadratic terms in x to the left-hand side and all the affine terms to the right-hand side, and $R(t)$ has the same meaning as before)

$$\begin{aligned}
&\frac{1}{2}\left|x(T)+\int_{t_0}^{T} G(T,s)x(s)\,ds+g(T)\right|_B^2+\frac{1}{2}\int_{t_0}^{T}\langle x(t),R(t)x(t)\rangle\,dt\\
&\quad+\int_{t_0}^{T}\int_{t_0}^{t}\langle x(t),(A^T(t)BG(t,s)-BC(t,s))x(s)\rangle\,ds\,dt\\
&\quad-\int_{t_0}^{T}\int_{t_0}^{t}\int_{t_0}^{t}\langle G(t,s)x(s),BC(t,v)x(v)\rangle\,dv\,ds\,dt\\
&=\frac{1}{2}|x_0+g(t_0)|_B^2+\int_{t_0}^{T}\langle f(t),Bg(t)\rangle\,dt\\
&\quad+\int_{t_0}^{T}\langle Bf(t)-A^T(t)Bg(t),x(t)\rangle\,dt\\
&\quad+\int_{t_0}^{T}\int_{t_0}^{t}\langle f(t),BG(t,s)x(s)\rangle\,ds\,dt\\
&\quad+\int_{t_0}^{T}\int_{t_0}^{t}\langle g(t),BC(t,s)x(s)\rangle\,ds\,dt.
\end{aligned} \tag{6.4}$$

Generally speaking, in this formula we can use the first two terms to dominate all the

rest. Most of the remaining perturbation terms are of a familiar nature; the only exception is the last term on the left-hand side (which can easily be estimated with Hölder's inequality).

We shall make no attempt to prove a general result based on (6.4), with all the terms present. There are simply too many different possible estimates. For example, in the last two terms, we may allow G and C to be "large" at infinity, if we instead require f and g to be "small," and vice versa (these terms can be estimated in a way similar to the one that we used at the end of the proof of Theorem 5.1). To prove [17, Thm. 1, p. 492] and [17, Thm. 5, p. 503] we drop the two functions f and g (after which the right-hand side of (6.4) is a constant), and estimate the last term on the left-hand side in the same way as Burton and Mahfoud do on the bottom of p. 491 in [17]. Here we shall only prove a result that applies to the equation which one gets by dropping the functions C and f on the right-hand side of (6.2), i.e., we look at the equation

$$\frac{d}{dt}\left(x(t)+\int_{t_0}^{t} G(t,s)x(s)\,ds+g(t)\right)+A(t)x(t)=0,\quad t\geqq t_0,\quad x(t_0)=x_0. \tag{6.5}$$

If we drop the corresponding terms in (6.4), and make the standard estimates, then we get

$$\begin{aligned}&\frac{1}{2}\left|x(T)+\int_{t_0}^{T} G(T,s)x(s)\,ds+g(T)\right|_B^2+\int_{t_0}^{T}\frac{p(t)}{2\beta}z^2(t)\,dt\\&\quad-\int_{t_0}^{T}\int_{t_0}^{t} z(t)\|B^{-1}A^T(t)BG(t,s)\|_B z(s)\,ds\,dt\\&\quad\leqq\frac{1}{2}|x_0+g(t_0)|_B^2+\int_{t_0}^{T}|A^T(t)Bg(t)||x(t)|\,dt,\qquad T\geqq t_0,\end{aligned} \tag{6.6}$$

and

$$\begin{aligned}&\frac{1}{2}\left|x(T)+\int_{t_0}^{T} G(T,s)x(s)\,ds+g(T)\right|_B^2+\int_{t_0}^{T}\frac{r(t)}{2}|x(t)|^2\,dt\\&\quad-\int_{t_0}^{T}\int_{t_0}^{t}|x(t)|\,\|A^T(t)BG(t,s)\|\,|x(s)|\,ds\,dt\\&\quad\leqq\frac{1}{2}|x_0+g(t_0)|_B^2+\int_{t_0}^{T}|A^T(t)Bg(t)||x(t)|\,dt,\qquad T\geqq t_0.\end{aligned} \tag{6.7}$$

These estimates combined with (5.4) gives us the following theorem.

THEOREM 6.1. (i) *If* $\int_{t_0}^{\infty}(1+p(t))^{-1}|A^T(t)Bg(t)|^2\,dt<\infty$, *and there are constants* $\kappa>0$ *and* $\varepsilon\in(0,1/2\beta)$ *such that*

$$\begin{aligned}&\frac{\kappa}{2}\int_{t_0}^{t}\|B^{-1}A^T(t)BG(t,u)\|_B\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|B^{-1}A^T(v)BG(v,t)\|_B\,dv\\&\quad\leqq\frac{p(t)}{2\beta}-\varepsilon(1+p(t))\end{aligned}$$

for almost all $t>t_0$, *or*

$$\begin{aligned}&\frac{\kappa}{2}\int_{t_0}^{t}\|A^T(t)BG(t,u)\|\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|A^T(v)BG(v,t)\|\,dv\\&\quad\leqq\frac{r(t)}{2}-\varepsilon(1+p(t))\end{aligned}$$

for almost all $t>t_0$, *then the solution* x *of* (6.5) *satisfies* $\int_{t_0}^{\infty}(1+p(t))|x(t)|^2\,dt<\infty$.

(ii) *In addition to* (i), *suppose that* $\int_{t_0}^{\infty}(1+p(t))^{-1}|g'(t)|^2\,dt<\infty$, *and that for some constant* M *and almost all* $t\geqq t_0$ *and* $s\geqq t_0$,

$$\|A(t)+G(t,t)\|\leqq M(1+p(t)),$$
$$\int_{t_0}^{t}\left\|\frac{\partial}{\partial t}G(t,s)\right\|ds\leqq M(1+p(t)),$$
$$\int_{s}^{\infty}\left\|\frac{\partial}{\partial t}G(t,s)\right\|dt\leqq M(1+p(s)).$$

Then $\int_{t_0}^{\infty}(|w'(t)|+(1+p(t))^{-1}|x'(t)|^2\,dt<\infty$, *where* $w(\cdot)=|x(\cdot)|_B^2$. *In particular,* $x(t)\to 0$ *as* $t\to\infty$.

The proof is essentially the same as the proof of parts (ii) and (iii) of Theorem 5.1.

Theorem 6.1 contains the linear versions of the stability claims in [16, Thm. 4, p. 150], [16, Thm. 14, p. 164], [16, Thm. 15, p. 166], and [16, Thm. 16, p. 167]. (The nonlinear versions are proved in an analogous way.)

7. The variation of constants formula. It is not possible to discuss equations (2.1), (2.6), and (6.2) without mentioning the most obvious approach to a stability theory for these equations, namely the use of the variation of constants formula. If we let $Z(t,s)$ denote the fundamental matrix solution of the equation $x'(t)+A(t)x(t)=0$, i.e., the solution of the equations

$$\frac{\partial}{\partial t}Z(t,s)+A(t)Z(t,s)=0,\qquad t_0\leqq s\leqq t<\infty,$$
$$\frac{\partial}{\partial s}Z(t,s)-Z(t,s)A(s)=0,\qquad t_0\leqq s\leqq t<\infty,$$
$$Z(t,t)=I,\qquad t\geqq t_0;$$

then we can multiply the equation

$$\frac{d}{dt}\left(x(t)+\int_0^t G(t,s)x(s)\,ds+g(t)\right)+A(t)x(t)=\int_{t_0}^{t}C(t,s)x(s)\,ds+f(t),\qquad t\geqq t_0,\quad x(t_0)=x_0,\tag{6.2}$$

by $Z(T,t)$, integrate over $[t_0,T]$, and finally integrate the term with the derivative on the left-hand side by parts to get the variation of constants formula

$$\begin{aligned}x(T)=(x_0+g(t_0))Z(T,t_0)-g(T)+\int_{t_0}^{T}Z(T,t)(f(t)+A(t)g(t))\,dt\\ -\int_{t_0}^{T}G(T,s)x(s)\,ds\\ +\int_{t_0}^{T}Z(T,t)\int_{t_0}^{t}(C(t,s)+A(t)G(t,s))x(s)\,ds\,dt.\end{aligned}$$

For notational convenience, let $\mathcal{B}$ denote one of the spaces $L^p([t_0,\infty);\mathbf{R}^n)$, $1\leqq p\leqq\infty$, $BUC([t_0,\infty);\mathbf{R}^n)$ (the space of bounded uniformly continuous functions on $[t_0,\infty)$), or $BC_0([t_0,\infty);\mathbf{R}^n)$ (the space of continuous functions tending to zero at infinity). Clearly, if all the operators $\mathcal{X}(x)(t)=\int_{t_0}^{t}Z(t,s)x(s)\,ds$, $\mathcal{G}(x)(t)=\int_{t_0}^{t}G(t,s)x(s)\,ds$, and $\mathcal{H}(x)(t)=\int_{t_0}^{t}(C(t,s)+A(t)G(t,s))x(s)\,ds$ are continuous from $\mathcal{B}$ into itself, if f and g belong to $\mathcal{B}$, and if the norm of the operators $\mathcal{G}$ and $\mathcal{H}$ are small enough, then we can use the contraction mapping principle to show that the solution of (6.2) belongs

to $\mathcal{B}$. It is even possible to let $\mathcal{G}$, $\mathcal{H}$, g, and f depend on x, as long as the dependence is so weak that the contraction mapping principle is applicable.

The approach described above is a minor modification of the approach used in [4]. Arguing as above one can remove the assumption in [4] that $A(t)$ commutes with $Z(t, s)$.

The following result is useful when one wants to show that the operators $\mathcal{L}$, $\mathcal{G}$, and $\mathcal{H}$ map some L^p-space into itself.

LEMMA 7.1. *Let $R \in L^1_{\text{loc}}([t_0, \infty) \times [t_0, \infty); \mathbf{R}^{n\times n})$, and denote the operator which maps x into the function $t \mapsto \int_{t_0}^t R(t, s)x(s)\, ds$ by $\mathcal{R}$. Then the following claims are true:*

(i) *$\mathcal{R}$ maps $L^\infty([t_0, \infty); \mathbf{R}^n)$ continuously into itself if and only if* $\operatorname{ess\,sup}_{t\geqq t_0} \int_{t_0}^t \|R(t, s)\|\, ds < \infty$.

(ii) *$\mathcal{R}$ maps $L^1([t_0, \infty); \mathbf{R}^n)$ continuously into itself if and only if* $\operatorname{ess\,sup}_{s\geqq t_0} \int_s^\infty \|R(t, s)\|\, dt \leqq \infty$.

(iii) *If $\mathcal{R}$ maps both $L^1([t_0, \infty); \mathbf{R}^n)$ and $L^\infty([t_0, \infty); \mathbf{R}^n)$ continuously into themselves, then $\mathcal{R}$ maps all intermediate spaces $L^p([t_0, \infty); \mathbf{R}^n)$, $1 < p < \infty$, continuously into themselves.*

All of these results are well known. No necessary and sufficient condition is known in the intermediate cases $1 < p < \infty$ (except for convolution kernels in the space L^2, which are continuous if and only if their distribution Fourier transforms are bounded). Additional admissibility results of this type are discussed in [18].

In particular, if there is some function $\gamma \in L^1([t_0, \infty), \mathbf{R})$ such that $\|R(t, s)\| \leqq \gamma(t-s)$, i.e., if R is dominated by a scalar L^1 convolution kernel, the $\mathcal{R}$ maps all the the spaces $L^p([t_0, \infty); \mathbf{R}^n)$, $1 \leqq p \leqq \infty$, continuously into themselves.

Instead of using the variation of constants formula for the equation $x'(t) + A(t)x(t) = 0$ one can use the variation of constants formula for (2.6). We let $R(t, s)$ be the differential resolvent of (2.6), i.e., the solution of the equations

$$
\begin{aligned}
&\frac{\partial}{\partial t} R(t, s) + A(t)R(t, s) = \int_s^t C(t, u)R(u, s)\, du \qquad t_0 \leqq s \leqq t < \infty,\\
&\frac{\partial}{\partial s} R(t, s) - R(t, s)A(s) = -\int_s^t R(t, u)C(u, s)\, du, \qquad t_0 \leqq s \leqq t < \infty,\\
&R(t, t) = I, \qquad t \geqq t_0,
\end{aligned} \tag{7.1}
$$

If we multiply (6.2) by $R(T, t)$, integrate over $[t_0, T]$, and finally integrate the term with the derivative on the left-hand side by parts, then we get

$$
\begin{aligned}
x(T) &= (x_0 + g(t_0))R(T, t_0) - g(T)\\
&\quad + \int_{t_0}^T R(T, t)(f(t) + A(t)g(t))\, dt - \int_{t_0}^T R(T, t) \int_{t_0}^t C(t, s)g(s)\, ds\, dt\\
&\quad - \int_{t_0}^T G(T, s)x(s)\, ds + \int_{t_0}^T R(T, t)A(t) \int_{t_0}^t G(t, s)x(s)\, ds\, dt\\
&\quad - \int_{t_0}^T R(T, t) \int_{t_0}^t C(t, u) \int_{t_0}^u G(u, s)x(s)\, ds\, du\, dt.
\end{aligned} \tag{7.2}
$$

This equation has a form that makes it possible to apply the contraction mapping principle under suitable smallness assumptions on G. In addition, one may let g, f and G depend on f (cf. [24]). We leave the exact formulation to the reader.

One point that ought to be made is that Theorems 3.1, 3.2, 3.3, 3.4, 4.2, 4.3, 5.1, 5.3, and 5.4 give sufficient conditions under which the differential resolvent maps

various spaces $\mathcal{B}$ continuously into themselves (to see this, take $x_0 = g = G = 0$ in (7.2)). Thus, they enable us to use the contraction mapping principle to study perturbations.

In the case where A is periodic with period T, and C satisfies $C(t+T, s+T) = C(t, s)$ for almost all s and t, then $R(t, s)$ can be defined for $-\infty < s \leqq t < \infty$, and, because of the fact that (under very weak conditions) the solution of (7.1) is unique, R automatically satisfies $R(t+T, s+T) = R(t, s)$ for almost all s and t. If, in addition, $\operatorname{ess\,sup}_{0 \leqq t \leqq T} \int_{-\infty}^{t} (\|C(t, s)\| + \|R(t, s)\|)\, ds < \infty$, or equivalently, $\operatorname{ess\,sup}_{t \geqq 0} \int_{0}^{t} (\|C(t, s)\| + \|R(t, s)\|)\, ds < \infty$, then it is not difficult to show (use Fubini's theorem and the resolvent equations (7.1)) that for each bounded periodic function f the equation

$$x'(t) + A(t) = \int_{-\infty}^{t} C(t, s)x(s)\, ds + f(t), \qquad -\infty < t < \infty, \tag{7.3}$$

has a unique bounded periodic solution, namely

$$x(t) = \int_{-\infty}^{t} R(t, s)f(s)\, ds, \qquad -\infty < t < \infty.$$

This means that under periodicity assumptions on A and C, if we can show that the solutions of (2.6) are bounded for bounded f, then we automatically get a periodicity result for (7.3). In the convolution case we may alternatively show that for each integrable f the solution of (2.6) is integrable, because this together with Lemma 7.1 will imply that solutions are bounded for bounded f (cf. the discussion of convolution equations given below).

8. Uniform stability and uniform asymptotic stability. When we formulate a stability result for a time dependent ordinary differential equation of the type

$$x'(t) + A(t)x(t) = 0, \quad t \geqq t_0, \quad x(t_0) = x_0, \tag{8.1}$$

we distinguish between the notions of stability, asymptotic stability, uniform stability, and uniform asymptotic stability. Stability at a point t_0 means that there is a constant $K(t_0)$, such that $|x(t)| \leqq K(t_0)|x(t_0)|$ for $t \geqq t_0$. The stability is uniform in an interval $J = [\beta, \infty)$, or in the interval $J = (-\infty, \infty)$, if the constant $K(t_0)$ can be made independent of $t_0 \in J$. The equation is asymptotically stable at t_0, if there is a bounded function γ_{t_0}, defined on $\mathbf{R}^+$ and tending to zero at infinity, such that $|x(t)| \leqq \gamma_{t_0}(t - t_0)|x(t_0)|$ for $t \geqq t_0$. Finally, it is uniformly asymptotically stable on an interval J if the function γ_{t_0} can be made independent of $t_0 \in J$.

In the special case where $C \equiv 0$ and $f \equiv 0$ in (2.6), it is obvious that the estimates that we have given above imply, under appropriate assumptions on the functions p and q, that (8.1) is stable, uniformly stable, or asymptotically stable on $\mathbf{R}$. A slightly less obvious fact is that they also can be used to prove uniform asymptotic stability on $\mathbf{R}$. For example, from Theorem 3.1(ii) it is easy to get a bound of the form $\|x\|_{L^1(t_0,\infty)} \leqq M|x(t_0)|$, for some constant M. Suppose that we have already proved that the equation is uniformly stable, i.e., suppose that $|x(t)| \leqq K|x(t_0)|$ for some K and all $t \geqq t_0$. Fix some $\varepsilon > 0$, and define $T(\varepsilon) = KM/\varepsilon$. Then, because of the upper bound that we have on $\|x\|_{L^1(t_0,\infty)}$, there must be at least one point t_1 in the interval $[t_0, t_0 + T(\varepsilon)]$ where $|x(t_1)| \leqq \varepsilon|x(t_0)|/K$. But this, together with the uniform stability, implies that $|x(t)| \leqq \varepsilon|x(t_0)|$ for $t \geqq t_1$. In particular, $|x(t)| \leqq \varepsilon|x(t_0)|$ for $t \geqq t_0 + T(\varepsilon)$. As $T(\varepsilon)$ is independent of t_0, this proves that the equation is uniformly asymptotically stable.

Maybe the main reason for why the property of being uniformly asymptotically stable is such an attractive one for an ordinary differential equation is the fact that uniform asymptotic stability implies exponential asymptotic stability, i.e., the function γ which we used above in the definition of uniform asymptotic stability can always be chosen so that it decays exponentially. (To see this, observe that if $\gamma(T) \leqq \frac{1}{2}$ for some T, then we can replace $\gamma(t)$ by $\frac{1}{2}\gamma(t-T)$ for $t \in [T, 2T]$, by $\frac{1}{4}\gamma(t-2T)$ for $t \in [2T, 3T]$, etc.) This means that the fundamental solution Z of (8.1) satisfies $\|Z(t, s)\| \leqq Me^{-\varepsilon(t-s)}$ for some positive constants M and ε. Therefore, the operator $\mathscr{L}$ defined in § 7 maps all L^p-spaces, $1 \leqq p \leqq \infty$, into themselves. As we saw in § 7, this property is extremely useful when one wants to prove perturbation results.

When we discuss the integral equation (2.6), we may very well interpret the notions of stability and asymptotic stability to mean exactly the type of results which we have given above. The notions of "uniform stability" and "uniform asymptotic stability" are much less clear. One interpretation of uniform stability would be to simply require the bounds that we get on x to be uniform in t_0. Clearly, such bounds follow immediately from our theorems.

Burton's interpretation of the notions uniform stability and uniform asymptotic stability is the following: He takes $f \equiv 0$, replaces the lower bound t_0 in the integral in (2.6) by zero, and specifies a bounded initial function φ on the interval $[0, t_0]$. The equation is supposed to hold for $t \geqq t_0$, and the solution is required to have the same type of uniform behavior as in the ODE case with $J = [t_0, \infty)$, but $|x(t_0)|$ is replaced by $\sup_{0 \leqq s \leqq t_0} |\varphi(s)|$.

To achieve a slightly greater generality, let us replace the interval $[0, t_0]$ by the interval $(-\infty, t_0]$, and specify an initial function on this interval. In other words, let us look at the initial value problem

$$\begin{aligned} &x'(t) + A(t)x(t) = \int_{-\infty}^{t} C(t, s)x(s)\, ds, \qquad t \geqq t_0, \\ &x(t) = \varphi(t), \qquad t \leqq t_0. \end{aligned} \tag{8.2}$$

(Clearly, we can consider (6.2) in a similar setting, but we leave this to the reader.) Let us follow Burton, and specify a bounded initial function (as opposed to an initial function in some L^p-space or weighted L^p-space). This equation is of the type (2.6), with f replaced by

$$f_{t_0,\varphi}(t) = \int_{-\infty}^{t_0} C(t, s)\varphi(s)\, ds, \tag{8.3}$$

and x_0 replaced by $\varphi(t_0)$. As

$$|f_{t_0,\varphi}(t)| \leqq \left(\int_{-\infty}^{t_0} \|C(t, s)\|\, ds \right) \sup_{s \leqq t_0} |\varphi(s)|, \tag{8.4}$$

and as the bounds that we have obtained on $\sup_{t \geqq t_0} |x(t)|$ throughout can be written as a constant times the maximum of $|x_0|$ and the appropriate L^p-norm of f, this means that it is very easy to get results on uniform stability in Burton's sense. The only thing that we have to do is to impose conditions on C that imply that the function $t \mapsto \int_{-\infty}^{t_0} \|C(t, s)\|\, ds$ has a uniformly bounded L^p-norm in the right L^p-space, so that the appropriate theorem applies. In particular, if we want to apply one of the results where $f \in L^1([t_0, \infty); \mathbf{R}^n)$, then it suffices to require that

$$\sup_{t_0 \in \mathbf{R}} \int_{t_0}^{\infty} \int_{-\infty}^{t_0} \|C(t, s)\|\, ds\, dt < \infty. \tag{8.5}$$

This is a very common condition (with the lower bound $-\infty$ replaced by 0) for uniform stability in [2]-[16]. The analogous condition for Theorem 5.1(ii) is

$$\sup_{t_0\in\mathbf{R}}\int_{t_0}^{\infty}(1+p(t))^{-1}\left[\int_{-\infty}^{t_0}\|C(t,s)\|\,ds\right]^2 dt<\infty. \tag{8.6}$$

For Theorem 4.2(ii) no special new condition is needed, because

$$(1+p(t))^{-1}\int_{-\infty}^{t_0}\|C(t,s)\|\,ds\leqq(1+p(t))^{-1}\int_{-\infty}^{t}\|C(t,s)\|\,ds$$

for $t\geqq t_0$.

The comments made above, together with Theorems 3.1(i), 4.2(i), 4.2(ii), 5.1(i), and 5.1(ii) prove the following theorem on uniform asymptotic stability in Burton's sense.

THEOREM 8.1. *Each of the following conditions imply that* (8.2) *is uniformly stable*:

(i) *Condition* (8.5) *holds, and* $\int_s^{\infty}\|C(t,s)\|_B\,dt\leqq p(s)/2\beta$ *for almost all* $s\in\mathbf{R}$, *or* $\int_s^{\infty}\|\sqrt{B}\,C(t,s)\|\,dt\leqq q(s)/2\sqrt{\beta}$ *for almost all* $s\in\mathbf{R}$.

(ii) *Condition* (8.5) *holds, and there is some* $\kappa>0$ *such that*

$$\frac{\kappa}{2}\int_{-\infty}^{t}\|C(t,u)\|_B\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|C(v,t)\|_B\,dv\leqq\frac{p(t)}{2\beta}$$

for almost all $t\in\mathbf{R}$, *or*

$$\frac{\kappa}{2}\int_{-\infty}^{t}\|BC(t,u)\|\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|BC(v,t)\|\,dv\leqq\frac{r(t)}{2}$$

for almost all $t\in\mathbf{R}$.

(iii) *Condition* (8.5) *holds, and* $\int_{-\infty}^{t}\|C(t,s)\|_B\,ds\leqq p(t)/2\beta$ *for almost all* $t\in\mathbf{R}$.

(iv) *Condition* (8.6) *holds, and there are constants* $\kappa>0$ *and* $\varepsilon\in(0,1/2\beta)$ *such that*

$$\frac{\kappa}{2}\int_{-\infty}^{t}\|C(t,u)_B\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|C(v,t)\|_B\,dv\leqq\frac{p(t)}{2\beta}-\varepsilon(1+p(t))$$

for almost all $t\in\mathbf{R}$, *or*

$$\frac{\kappa}{2}\int_{-\infty}^{t}\|BC(t,u)\|\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|BC(v,t)\|\,dv\leqq\frac{r(t)}{2}-\varepsilon(1+p(t))$$

for almost all $t\in\mathbf{R}$.

(v) $\int_{-\infty}^{t}\|C(t,s)\|_B\,ds\leqq p(t)/2\beta-\varepsilon(1+p(t))$ *for some constant* $\varepsilon\in(0,1/2\beta)$ *and almost all* $t\in\mathbf{R}$.

We leave it to the reader to formulate similar results based on the remaining theorems in §§ 3-6.

The question of uniform asymptotic stability is more delicate, and concrete results on uniform asymptotic stability are scarce in [2]-[17] (here the word "concrete" refers to results that have been applied to (2.6) or (6.2)). Moreover, it is not clear to what extent this property implies that the operator $\mathcal{R}$ induced by the resolvent R (cf. § 7) maps various spaces of functions into themselves, as it does for ordinary differential equations. The convolution case in an exception; see § 8.

The main result in [10] on the uniform asymptotic stability of (2.6) (in the case when (2.6) is not a convolution equation) seems to be [10, Thm. 2.5.1(d), p. 39]. That theorem is closely related to our Theorem 3.1, so it is to be expected that one should be able to prove a version of Theorem 3.1 where one has a uniform rate of convergence to zero. This is indeed the case.

First, let us prove an auxiliary lemma.

LEMMA 8.2. *Let* $x \in L^1([t_0, \infty); \mathbf{R}^n)$, *and* $x' \in L^1([t_0, \infty); \mathbf{R}^n)$. *For each* $\varepsilon > 0$, *define* $S(\varepsilon) = 2\|x\|_{L^1(t_0,\infty)}/\varepsilon$ *and* $N(\varepsilon) = 2\|x'\|_{L^1(t_0,\infty)}/\varepsilon + 1$. *Then, for every* $T \geqq S(\varepsilon)$, *the interval* $[t_0, t_0 + N(\varepsilon)T]$ *contains at least one subinterval of length* T *on which* $|x(t)| \leqq \varepsilon$.

The proof given below has been adopted from [10, pp. 40–41].

Proof. Clearly, each interval of length T contains at least one point t_1 where $|x(t_1)| \leqq \frac{1}{2}\varepsilon$ (otherwise the L^1-norm of x over this interval is too big). If it also contains a point t_2 where $|x(t_2)| \geqq \varepsilon$, then the integral of $|x'|$ over this interval is at least $\frac{1}{2}\varepsilon$. Because of the bound that we have on the L^1-norm of x', there can be at most $2/\varepsilon\|x'\|_{L^1(t,\infty)}$ (rounded upwards to the nearest integer) intervals of this type, containing points of both the types t_1 and t_2. From this the conclusion follows. □

Suppose that (8.5) holds, and that the assumption on C in Theorem 3.1(ii) is satisfied, with a constant ε independent of t_0. Let x be the solution of (8.2). Then, from Theorem 3.1(ii) we get a bound of the form

$$\sup_{t \geqq t_0} |x(t)| + \|x\|_{L^1(t_0,\infty)} + \|x'\|_{L^1(t_0,\infty)} \leqq M(|\varphi(t_0)| + \|f_{t_0,\varphi}\|_{L^1(t_0,\infty)}), \tag{8.7}$$

where the constant M is independent of t_0 and φ. But (8.4) and (8.5) imply that

$$\|f_{t_0,\varphi}\|_{L^1(t_0,\infty)} \leqq K \sup_{s \leqq t_0} |\varphi(s)|, \tag{8.8}$$

where K is the supremum on the left-hand side of (8.5). Therefore,

$$\sup_{t \geqq t_0} |x(t)| + \|x\|_{L^1(t_0,\infty)} + \|x'\|_{L^1(t_0,\infty)} \leqq M(1+K) \sup_{s \leqq t_0} |\varphi(s)|. \tag{8.9}$$

By (8.9) and Lemma 8.2 (with ε replaced by $\varepsilon M(1+K)\sup_{s\leqq t_0}|\varphi(s)|$), for every $\varepsilon > 0$ and every $T \geqq 2/\varepsilon$, we can within the interval $[t_0, t_0 + (2/\varepsilon + 1)T]$ find a subinterval of length T on which $|x(t)| \leqq \varepsilon M(1+K)\sup_{s\leqq t_0}|\varphi(s)|$. If we want this fact to imply that (8.2) is uniformly asymptotically stable, then we need the following property of the function $f_{t_0,\varphi}$:

(8.10) There exists a constant K and a function $V(\varepsilon)$ such that for all $\varepsilon > 0$, all $t_0 \in \mathbf{R}$, and all bounded φ, it is true that $\|f_{t_0,\varphi}\|_{L^1(t_0,\infty)} \leqq K(\sup_{t_0 - V(\varepsilon) \leqq s \leqq t_0} |\varphi(s)| + \varepsilon \sup_{s \leqq t_0 - V(\varepsilon)} |\varphi(s)|)$.

When this property is satisfied, we get uniform asymptotic stability by the following argument: Fix some $\varepsilon > 0$. Define $T = \max\{2/\varepsilon, V(\varepsilon)\}$. Let t_1 be the right end point of the interval referred to above where $|x(t)| \leqq \varepsilon M(1+K)\sup_{s\leqq t_0}|\varphi(s)|$. Then $t_1 - t_0 \leqq (2/\varepsilon + 1)T$ (so we get an upper bound on $t_1 - t_0$ that depends on ε, but is independent of t_0 and φ). For $t \in [t_0, t_1]$, define $\varphi(t) = x(t)$. Then $\sup_{s\leqq t_1 - V(\varepsilon)}|\varphi(s)\| \leqq M(1+K)\sup_{s\leqq t_0}|\varphi(s)|$ and $\sup_{t_1 - V(\varepsilon) \leqq s \leqq t_1}|\varphi(s)| \leqq \varepsilon M(1+K)\sup_{s\leqq t_0}|\varphi(s)|$. Therefore, by (8.10), we have $\|f_{t,\varphi}\|_{L^1(t_1,\infty)} \leqq 2\varepsilon MK(1+K)\sup_{s\leqq t_0}|\varphi(s)|$. However, by (8.7), this means that for $t \geqq t_1$, we have $|x(t)| \leqq \varepsilon M^2(1+K)(2+K)\sup_{s\leqq t_0}|\varphi(s)|$. Thus, the solution tends to zero with a convergence rate that is independent of t_0. (The reader who prefers to end up with a clean result should divide ε by $M^2(1+K)(2+K)$ at the beginning.)

It only remains to transfer (8.10) into an assumption on C. By (8.3),

$$|f_{t_0,\varphi}| \leqq \left(\int_{-\infty}^{t_0 - V(\varepsilon)} \|C(t,s)\|\, ds\right) \sup_{s \leqq t_0 - V(\varepsilon)} |\varphi(s)| + \left(\int_{t_0 - V(\varepsilon)}^{t_0} \|C(t,s)\|\, ds\right) \sup_{t_0 - V(\varepsilon) \leqq s \leqq t_0} |\varphi(s)|$$

and therefore, it suffices to assume, in addition to (8.5), that

$$\lim_{T\to\infty}\sup_{t_0\in\mathbf{R}}\int_{t_0}^{\infty}\int_{-\infty}^{t_0-T}\|C(t,s)\|\,ds\,dt=0. \tag{8.11}$$

(This is the same condition which Burton uses in [10, Thm. 2.5.1(d)]. It says that the kernel "forgets" old values of the function φ with a uniform rate in a certain sense.)

The argument above proves the following theorem.

THEOREM 8.3. *Let* (8.5) *and* (8.11) *hold. In addition, suppose that for some constant* $\varepsilon\in(0,1/2\beta)$, *it is true that* $\int_s^{\infty}\|C(t,s)\|_B\,dt\leqq p(s)/2\beta-\varepsilon(1+p(s))$ *for almost all* $s\in\mathbf{R}$, *or that* $\int_s^{\infty}\|\sqrt{B}\,C(t,s)\|\,dt\leqq q(s)/2\sqrt{\beta}-\varepsilon(1+p(s))$ *for almost all* $s\in\mathbf{R}$. *Then* (8.2) *is uniformly asymptotically stable on* $\mathbf{R}$.

The same argument, but with Theorem 3.1(ii) replaced by Theorem 5.1(ii), gives the following result.

THEOREM 8.4. *Suppose that there are constants* $\kappa>0$ *and* $\varepsilon\in(0,1/2\beta)$ *such that*

$$\frac{\kappa}{2}\int_{-\infty}^{t}\|C(t,u)\|_B\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|C(v,t)\|_B\,dv\leqq\frac{p(t)}{2\beta}-\varepsilon(1+p(t))$$

for almost all $t\in\mathbf{R}$, *or*

$$\frac{\kappa}{2}\int_{-\infty}^{t}\|BC(t,u)\|\,du+\frac{1}{2\kappa}\int_{t}^{\infty}\|BC(v,t)\|\,dv\leqq\frac{r(t)}{2}-\varepsilon(1+p(t))$$

for almost all $t\in\mathbf{R}$. *In addition, suppose that* (8.6) *holds, and that*

$$\lim_{T\to\infty}\sup_{t_0\in\mathbf{R}}\int_{t_0}^{\infty}(1+p(t))^{-1}\left[\int_{-\infty}^{t_0-T}\|C(t,s)\|\,ds\right]^2dt=0. \tag{8.12}$$

Then (8.2) *is uniformly asymptotically stable on* $\mathbf{R}$.

(To prove this theorem, argue as in the proof of Theorem 8.3, but apply Lemma 8.2 to the function $|x|^2$ instead of to the function x.)

This result seems to be new.

The analogue of (8.11) and (8.12) for Theorem 4.2(ii) would be

$$\lim_{T\to\infty}\sup_{t_0\in\mathbf{R}}\operatorname*{ess\,sup}_{t\geqq t_0}(1+p(t))^{-1}\int_{-\infty}^{t_0-T}\|C(t,s)\|\,ds=0.$$

However, this condition is not enough for uniform asymptotic stability, because Lemma 8.2 no longer applies. It is an interesting open problem, under what conditions one gets uniform asymptotic stability in Theorem 4.2.

9. Convolution equations. In our preceding discussions we have ignored all the available Lyapunov results for convolution equations. The convolution versions of (2.6) and (7.1) are

$$x'(t)+Ax(t)=\int_0^t C(t-s)x(s)\,ds+f(t),\quad t\geqq 0,\quad x(0)=x_0 \tag{9.1}$$

and

$$\begin{aligned}&R'(t)+A(t)R(t)=\int_0^t C(t-s)R(s)\,du, && t\geqq 0,\\ &R'(t)+R(t)A(t)=\int_0^t R(t-s)C(s)\,du, && t\geqq 0,\\ &R(0)=I.\end{aligned} \tag{9.2}$$

For linear convolution equations all significant stability properties can be related to one of the three conditions:

(1) The differential resolvent R is bounded;

(2) The differential resolvent R tend to zero at infinity;

(3) The differential resolvent R is integrable.

Especially the third condition is of fundamental importance; it is equivalent both to uniform stability and to uniform asymptotic stability. A necessary and sufficient condition is known for R to be integrable in the case where $C \in L^1(\mathbf{R}^+; \mathbf{R}^{n\times n})$: The Laplace transform $\hat{C}$ of C must satsify $\det(zI + A - \hat{C}(z)) \neq 0$ for $\Re z \geqq 0$ (see [24, Thm. 3.5, p. 558] for the scalar case). The same Laplace transform condition is necessary and sufficient for the integrability of the differential resolvent also, e.g., in the cases where $\hat{C}(0) = \lim_{t\to\infty} \int_0^t C(s)\,ds$ exists, and $\int_0^\infty \| \int_0^t C(s)\,ds - \hat{C}(0)\|\,dt < \infty$, and where C is of bounded variation and $C(\infty)$ is invertible (this follows from [30, Prop. 2.3, p. 755]). A similar result is true for the convolution version of (6.2), namely

$$\frac{d}{dt}\left(x(t) + \int_0^t G(t-s)x(s)\,ds + g(t)\right) + Ax(t) = \int_0^t C(t-s)x(s)\,ds + f(t), \tag{9.3}$$
$$t \geqq 0, \quad x(0) = x_0.$$

These results are quite powerful, and they supersede the Lyapunov theory for these classes of equations in the sense that in all the cases where one has used Lyapunov methods to prove uniform stability or uniform asymptotic stability, it is possible to check that the Laplace transform condition is satisfied. This applies, in particular, to [4, Cor., p. 396], [8, Thm. 6, p. 284], [11, Prop., p. 242], [16, Thm. 1, p. 145], [16, Thm. 2, p. 146], [16, Thm. 3′, p. 152], [16, Thm. 4′, p. 152], [16, Thm. 5, p. 153], [16, Cor. 3, p. 158], [16, Thm. 10, p. 159], [16, Thm. 11, p. 161], [16, Thm. 12, p. 162], [17, Thm. 4, p. 498], and [17, Thm. 9, p. 511].

(To show that the Laplace transform condition is satisfied in the convolution versions of Theorems 3.1(ii), 4.2(ii), 5.1(ii), and 5.3(ii), we can use the following simple observation: Define $\Delta(z) = zI + A - \hat{C}(z)$. If $\det \Delta(z_0) = 0$ for some z_0 with $\Re z_0 \geqq 0$, then there is some nonzero $w \in \mathbf{C}^n$ for which $\Delta(z_0)w = 0$; hence $\langle w, B\Delta(z_0)w\rangle = 0$. Thus, to prove that $\det \Delta(z) \neq 0$ for $\Re z \geqq 0$, it certainly suffices to show that for all nonzero $w \in \mathbf{R}^n$, we have $\Re\langle w, B\Delta(z)w\rangle > 0$ for $\Re z \geqq 0$. If we denote $A^TB + BA$ by R, then this is equivalent to showing that $|w|_B^2 \Re z + \frac{1}{2}\langle w, Rw\rangle > \Re\langle w, B\hat{C}(z)w\rangle$ for all nonzero $w \in \mathbf{C}^n$ and all $z \in \mathbf{C}$ with $\Re z \geqq 0$. In particular, it suffices to show that $|w|_B^2 \Re z + \frac{1}{2}\langle w, Rw\rangle > \int_0^\infty |\langle w, BC(t)w\rangle|\,dt$, because $\Re\langle w, B\hat{C}(z)w\rangle \leqq |\langle w, B\hat{C}(z)w\rangle| \leqq \int_0^\infty |\langle w, BC(t)w\rangle|\,dt$ for such w and z.)

Results on specific decay rates of linear convolution equations are available through Laplace transform methods as well, see e.g. [30] and [37, Thm. 3, p. 318] (some special cases of this theorem were proved independently in [14, Thm. 4, p. 660] and [14, Thm. 5, p. 661]).

Another class of equations which we have ignored above is the class of nonlinear equations of the type

$$x'(t) + AG(x(t)) = \int_0^t C(t-s)G(x(s))\,ds + f(t), \quad t \geqq 0, \quad x(0) = x_0, \tag{9.4}$$

with a kernel of positive type. The traditional approach is to study these equations by means of Lyapunov methods (see Levin's classical papers [31] and [32]), but for this

class of equations the Lyapunov method has been superseded by the energy method (see, e.g., [38], [39], and [40]).

The question of specific decay rates for solutions of a scalar version of (9.4) have been studied in [42]. The proofs given there are based on an L^2-technique. Similar results are proved in [14] with a Lyapunov technique.

REFERENCES

[1] S. R. BERNFELD AND J. R. HADDOCK, *Liapunov-Razumikhin functions and convergence of solutions of functional differential equations*, Applicable Anal., 9 (1979), pp. 235-245.

[2] T. A. BURTON, *Stability theory for Volterra equations*, J. Differential Equations, 32 (1979), pp. 101-118.

[3] ———, *Uniform stabilities for Volterra equations*, J. Differential Equations, 36 (1980), pp. 40-53.

[4] ———, *An integrodifferential equation*, Proc. Amer. Math. Soc., 79 (1980), pp. 393-399.

[5] ———, *Construction of Liapunov functionals for Volterra equations*, J. Math. Anal. Appl., 85 (1982), pp. 90-105.

[6] ———, *Perturbed Volterra equations*, J. Differential Equations, 43 (1982), pp. 168-183.

[7] ———, *Boundedness in functional differential equations*, Funkcial. Ekvac., 25 (1982), pp. 51-77.

[8] ———, *Volterra equations with small kernels*, J. Integral Equations, 5 (1983), pp. 271-285.

[9] ———, *Structure of solutions of Volterra equations*, SIAM Rev. 25 (1983), pp. 343-364.

[10] ———, *Volterra Integral and Differential Equations*, Academic Press, New York, London, 1983.

[11] ———, *Periodic solutions of linear Volterra equations*, Funkcial. Ekvac., 27 (1984), pp. 229-253.

[12] ———, *Phase space and boundedness in Volterra equations*. J. Integral Equations, 10 (1985), pp. 61-72.

[13] ———, *Stability and Periodic Solutions for Ordinary and Functional Differential Equations*, Academic Press, New York, London, 1985.

[14] T. A. BURTON, Q. HUANG, AND W. E. MAHFOUD, *Rate of decay of solutions of Volterra equations*, Nonlinear Anal., 9 (1985), pp. 651-663.

[15] ———, *Liapunov functionals of convolution type*, J. Math. Anal. Appl., 106 (1985), pp. 249-272.

[16] T. A. BURTON AND W. E. MAHFOUD, *Stability criteria for Volterra equations*, Trans. Amer. Math. Soc., 279 (1983), pp. 143-174.

[17] ———, *Stability by decomposition for Volterra equations*, Tôhoku Math. J., 37 (1985), pp. 489-511.

[18] C. CORDUNEANU, *Integral Equations and Stability of Feedback Systems*, Academic Press, New York, London, 1973.

[19] H. ENGLER, *Bounds and asymptotics for a scalar Volterra integral equation*, J. Integral Equations, 7 (1984), pp. 209-227.

[20] ———, *On nonlinear scalar Volterra integral equations*, I, Trans. Amer. Math. Soc., 291 (1985), pp. 319-336.

[21] ———, *A note on scalar Volterra integral equations*, II, J. Math. Anal. Appl., 115 (1986), pp. 363-395.

[22] R. GRIMMER AND G. SEIFERT, *Stability properties of Volterra integrodifferential equations*, J. Differential Equations, 19 (1975), pp. 142-166.

[23] G. GRIPENBERG, *On some positive definite forms and Volterra integral operators*, Applicable Anal., 11 (1981), pp. 211-222.

[24] S. I. GROSSMAN AND R. K. MILLER, *Nonlinear Volterra integrodifferential systems with L^1-kernels*, J. Differential Equations, 13 (1973), pp. 551-566.

[25] J. R. HADDOCK, *Some new results on stability and convergence of solutions of ordinary and functional differential equations*, Funkcial. Ekvac., 19 (1976), pp. 247-269.

[26] J. R. HADDOCK AND J. TERJÉKI, *Liapunov-Razumikhin functions and an invariance principle for functional differential equations*, J. Differential Equations, 48 (1983), pp. 95-122.

[27] J. R. HADDOCK AND T. KRISZTIN, *Estimates regarding the decay of solutions of functional differential equations*, Nonlinear Anal., 8 (1984), pp. 1395-1408.

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

[29] ———, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, New York, 1975.

[30] G. S. JORDAN, O. J. STAFFANS, AND R. L. WHEELER, *Local analyticity in weighted L^1-spaces and applications to stability problems for Volterra equations*, Trans. Amer. Math. Soc., 274 (1982), pp. 749-782.

[31] J. J. LEVIN, *The asymptotic behavior of the solution of a Volterra equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 534-541.

[32] ———, *The qualitative behavior of a nonlinear Volterra equation*, Proc. Amer. Math. Soc., 16 (1965), pp. 711-718.

[33] J. J. LEVIN, *On a nonlinear Volterra equation*, J. Math. Anal. Appl., 39 (1972), pp. 458-476.
[24] ———, *A bound on the solutions of a Volterra equation*, Arch. Rational Mech. Anal., 52 (1973), pp. 339-349.
[35] D. L. LUKES, *Differential Equations: Classical to Controlled*, Academic Press, New York, London, 1982.
[36] R. K. MILLER, *Nonlinear Volterra Integral Equations*, Benjamin, Menlo Park, CA, 1971.
[37] D. F. SHEA AND S. WAINGER, *Variants of the Wiener-Lévy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1972), pp. 312-343.
[38] O. J. STAFFANS, *Positive definite measures with applications to a Volterra equation*, Trans. Amer. Math. Soc., 218 (1976), pp. 219-237.
[39] ———, *Tauberian theorems for a positive definite form, with applications to a Volterra equation*, Trans. Amer. Math. Soc., 218 (1976), pp. 239-259.
[40] ———, *Systems of nonlinear Volterra equations with positive definite kernels*, Trans. Amer. Math. Soc., 228 (1977), pp. 99-116.
[41] ———, *Boundedness and asymptotic behavior of solutions of a Volterra equation*, Michigan Math. J., 24 (1977), pp. 77-95.
[42] ———, *A nonlinear Volterra equation with rapidly decaying solutions*, Trans. Amer. Math. Soc., 258 (1980), pp. 523-530.
[43] ———, *A bound on the solutions of a nonlinear Volterra equation*, J. Math. Anal. Appl., 83 (1981), pp. 127-134.
[44] ———, *A neutral FDE with a stable D-operator is retarded*, J. Differential Equations, 49 (1983), pp. 208-217.

SIAM J. MATH ANAL.
Vol. 19, No. 4, July 1988

010

RIEMANN FUNCTION OF HARMONIC EQUATION AND APPELL'S F_4*

KATSUNORI IWASAKI†

Abstract. In this paper we calculate the Riemann function of Darboux's Harmonic equation (H). The Riemann function of (H) admits an action of a certain finite group, which permits us to reduce (H) to the system of Appell's F_4. Then, by using Takano's integral representation for F_4, we solve a connection problem to find a representation of R in terms of the Appell's hypergeometric function. We also discuss the symmetry algebras for classical and typical hyperbolic partial differential equations.

Key words. Riemann function, Harmonic equation, Euler-Poisson-Darboux equation, Appell's F_4, symmetry algebra, integral representation, connection problem

AMS(MOS) subject classifications. 33A35, 35C05, L10Q05

1. Introduction. In general, the solution space of a linear partial differential equation is not finite-dimensional. However, it can have a "fundamental solution" represented in terms of a solution of an ordinary differential equation or a system of partial differential equations with finite-dimensional solution space. In a sense, such a partial differential equation may be considered as one of good nature.

A few examples are the Euler-Poisson-Darboux (EPD), Confluent Euler-Poisson-Darboux (CEPD) and Telegraphic (TEL) equations:

$$\text{(EPD)} \qquad \{(x-y)(\partial^2/\partial x\partial y)+a(\partial/\partial x)-b(\partial/\partial y)\}u=0,$$

$$\text{(CEPD)} \qquad \{x(\partial^2/\partial x\partial y)+(\partial/\partial x)-b(\partial/\partial y)\}u=0,$$

$$\text{(TEL)} \qquad \{(\partial^2/\partial x\partial y)-1\}u=0,$$

where a and b are complex numbers. In a special case where $a=b=0$, (EPD) reduces to the Wave equation (W). Equation (CEPD) is obtained from (EPD) by a confluent procedure, and (TEL) from (CEPD). It is known as Appell's theorem (Darboux [2], Miller [7]) that the Lie group $SL_2(\mathbb{C})$ acts on the solution space of (EPD) by

$$u(x,y)\mapsto(\gamma x+\delta)^a(\gamma y+\delta)^b u\left(\frac{\alpha x+\beta}{\gamma x+\delta},\frac{\alpha y+\beta}{\gamma y+\delta}\right), \qquad \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}\in SL(2,\mathbb{C}).$$

By using this fact, we can show that a fundamental solution (the Riemann function) of (EPD) can be represented in terms of the Gauss' hypergeometric function (see Darboux [2], and, in a special case where $a=b$, see Iwanami Sûgaku Ziten [4]). Then, by a confluence, the Riemann functions of (CEPD) and (TEL) turn out to be represented in terms of the confluent hypergeometric function and the Bessel function, respectively. We note that (CEPD) and (TEL) also admit actions of certain Lie groups, which are alternatively used for a determination of their Riemann functions. These examples show that a situation in which a Lie group (or a Lie algebra) acts on the solution space of a differential equation is very important.

We consider a partial differential equation of the form

$$(1) \qquad Du=\{a(\partial^2/\partial x\partial y)+b(\partial/\partial x)+c(\partial/\partial y)+d\}u=0,$$

a, b, c, d: functions of x and y,

* Received by the editors February 17, 1987; accepted for publication (in revised form) July 21, 1987.
† Department of Mathematics, Faculty of Science, University of Tokyo, Hongô, Tokyo 113, Japan.

which contains the above examples as special cases. For the reason mentioned above, the symmetry algebra of (1) (see Miller [6], [7]) may give us some information about its Riemann function. If (1) contains parameters, then the parameter changing symmetry algebra (Miller [7]) may also be useful. In fact, we will prove (Theorem 1) that, if the dimension of the *reduced* symmetry algebra (see § 2 for definition) is bigger than one, then (1) is essentially equivalent to either (EPD), (CEPD), or (TEL). Hence the former algebra is useless for all equations except for the above three examples.

In this paper, we shall consider another example

$$\text{(H)} \qquad L_{(x,y)}u = \left\{\frac{\partial^2}{\partial x \partial y} + \frac{a(a-1)}{(x-y)^2} - \frac{b(b-1)}{(x+y)^2}\right\}u = 0, \qquad a, b \in \mathbb{C},$$

which is called *équations harmoniques* by Darboux [2]. The symmetry algebra of (H) is trivial and not very useful, but its parameter changing symmetry algebra becomes $sl_4(\mathbb{C})$, (§ 2), which will reveal a relation between the Riemann function of (H) and Appell's F_4. In this paper, however, we take another approach. We find that a finite group isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ acting on its Riemann function permits us to reduce (H) to Appell's system (F_4) (§§ 4–6). Then, by using an integral representation for F_4 obtained by Takano [10], we solve a connection problem to find a representation of the Riemann function in terms of Appell's hypergeometric function F_4 (§§ 7–10). A final result is given in Theorem 20 (§ 11).

We note that (H) arises from the complex wave equation $v_{tt} - \Delta_3 v = 0$ in four-dimensional space time, which has $sl_4(\mathbb{C})$-symmetry, by separating two angular coordinates. The parameters a and b of (H) are simply separation constants. Hence we can also examine (H) from the point of view of Kalnins and Miller [8].

Recall that the Riemann function $R(x, y; \xi, \eta)$ of (H) is a function of four variables satisfying the condition

$$LR = 0 \qquad \text{(as a function of } x \text{ and } y\text{)}, \tag{2}$$

$$(x-\xi)(y-\eta) = 0 \text{ implies } R = 1. \tag{3}$$

In this paper, we use the notation $\partial_x = \partial/\partial x$, $\delta_x = x(\partial/\partial x)$, etc.

2. Symmetry algebra. Following Miller [6], we define the *symmetry algebra* $\mathcal{S}$ of (1) by a Lie algebra of operators S of the form

$$S = \varphi \partial_x + \phi \partial_y + \psi, \qquad \varphi, \phi, \psi: \text{ functions of } x \text{ and } y,$$

$$[D, S] \equiv 0 \pmod{D}.$$

The Lie algebra $\mathcal{S}$ contains a trivial ideal $\mathbb{C}$. We call the quotient algebra $\mathcal{S}/\mathbb{C}$ the *reduced symmetry algebra* and denote it by $\hat{\mathcal{S}}$. If (1) has polynomial coefficients, then, from a practical point of view, it is convenient to restrict symmetry operators S to those for which φ, ϕ and ψ are polynomials. Thus we define the (reduced) polynomial symmetry algebra $\mathcal{S}_{pol}$ ($\hat{\mathcal{S}}_{pol}$) in a similar manner. For the above examples, the symmetry algebras are given, respectively, by (Miller [7], Okamoto [9])

$$\hat{\mathcal{S}}_{pol} = \{H, X, Y, 1\} \quad \text{(generators)},$$

where

$$\text{(EPD)} \qquad \begin{array}{lll} H = -2\delta_x - 2\delta_y - a - b, & [X, Y] = H, & \\ X = -\delta_x - \delta_y, & [H, X] = 2X, & \mathcal{S}_{pol} = sl_2(\mathbb{C}) \oplus \mathbb{C}, \\ Y = x\delta_x + y\delta_y + bx + ay, & [H, Y] = -2Y, & \hat{\mathcal{S}}_{pol} = sl_2(\mathbb{C}), \end{array}$$

$$\text{(CEPD)}\quad \begin{aligned} &H=-2\delta_x-2\delta_y, && [X,Y]=-1, && \mathcal{S}_{pol}=T_3\vec{\oplus}\mathbb{C},\\ &X=\partial_y, && [H,X]=2X, && \hat{\mathcal{S}}_{pol}=T_3,\\ &Y=x\delta_x-bx-y, && [H,Y]=-2Y, \end{aligned}$$

$$\text{(TEL)}\quad \begin{aligned} &H=-2\delta_x+2\delta_y, && [X,Y]=0, && \mathcal{S}_{pol}=T_3\oplus\mathbb{C},\\ &X=\partial_x, && [H,X]=2X, && \hat{\mathcal{S}}_{pol}=T_3,\\ &Y=\partial_y, && [H,Y]=-2Y, \end{aligned}$$

$$\text{(W)}\qquad \mathcal{S}=\mathcal{W}_x\oplus\mathcal{W}_y\oplus\mathbb{C},\qquad \hat{\mathcal{S}}=\mathcal{W}_x\oplus\mathcal{W}_y,$$

where $\oplus$ denotes the direct sum of Lie algebras, and $T_3\vec{\oplus}\mathbb{C}$ stands for a nontrivial central extension of T_3 by $\mathbb{C}$. Moreover, $\mathcal{W}_x$ is the Lie algebra of all functions in one variable x with Lie bracket $[f,g]=f(\partial_x g)-(\partial_x f)g$, the Wronskian of f and $g\in\mathcal{W}$.

THEOREM 1. *If the dimension of the reduced symmetry algebra $\hat{\mathcal{S}}$ is bigger than one, then* (1) *is equivalent to one of* (EPD), (CEPD), *or* (TEL) *up to transformations of independent variables $(x,y)\mapsto(X,Y)$ and those of dependent variable $u\mapsto U$ of the form*:

$$X=f(x),\quad Y=g(y),\quad U=h(x,y)u,$$

f, g, h: functions of x, y and (x,y), respectively.

We omit a proof of this theorem. Okamoto [9] independently showed a similar result in a different situation. Only we point out that there is an injective homomorphism from $\hat{\mathcal{S}}$ to $\mathcal{W}_x\oplus\mathcal{W}_y$

$$\hat{\mathcal{S}}\ni S\ (\mathrm{mod}\ \mathbb{C})\mapsto(\varphi,\phi)\in\mathcal{W}_x\oplus\mathcal{W}_y,$$

and that (1) admits an infinite-dimensional symmetry algebra only if (1) = (W). Therefore, in view of Lemma 2 below, nontrivial symmetry algebras that can act on (1) are very restricted, where we say "nontrivial" if $\dim\hat{\mathcal{S}}>1$ and "trivial" if $\dim\hat{\mathcal{S}}\leqq 1$.

LEMMA 2. *Finite-dimensional subalgebra of $\mathcal{W}$ is one-, two-, or three-dimensional and, in the latter two cases, it is noncommutative. Hence there are just three cases up to isomorphisms of Lie algebra.*

A proof of this lemma will be given in the Appendix (§ 12).

From now on, we consider the (polynomial) symmetry algebra $\mathcal{S}_{pol}$ for $D=(x-y)^2(x+y)^2L$, which we call the symmetry algebra of (H). We separate (H) to a *special case* and a *generic case* according to whether $a(a-1)$ is equal to $b(b-1)$ or not. Compared with those of (EPD), (CEPD), and (TEL), the symmetry algebra of (H) is trivial in a generic case. In a special case it is of some interest, but then (H) is essentially reduced to (EPD).

PROPOSITION 3. (i) *In a generic case, $\mathcal{S}_{pol}$ has a single generator $\delta_x+\delta_y$.*

(ii) *In a special case, $\mathcal{S}_{pol}$ is generated by $\delta_x+\delta_y$ and $x^2\delta_x+y^2\delta_y$.*

In fact, in a generic case, we can show that $\mathcal{S}=\mathcal{S}_{pol}$. In view of Theorem 1 and Proposition 3 (ii), in a special case (H) can be reduced to one of (EPD), (CEPD), or (TEL). Indeed, we have the following.

PROPOSITION 4. *In a special case, the substitution $X=x^2$, $Y=y^2$ converts* (H) *to*

$$\text{(SEPD)}\qquad \{(\partial^2/\partial X\partial Y)+a(a-1)/(X-Y)^2\}u=0\quad (\textit{special}\ \text{EPD}).$$

Moreover, the substitution $u=(X-Y)^a v$ reduces (SEPD) *to* (EPD) *with $a=b$. The fractional linear (Möbius) transformation group* Möb *acts on* (SEPD) *by*

$$u(X,Y)\mapsto u(AX,AY),\qquad A\in\text{Möb}. \tag{4}$$

We shall here refer to the parameter changing symmetry algebra $\mathcal{T}$ of (H). It is a Lie algebra of operators T of the form

$$T = \varphi\partial_x + \phi\partial_y + \xi\partial_s + \eta\partial_t + \psi, \qquad [L', T] \equiv 0 \pmod{L'},$$

where L' is an operator defined by

$$L' = (x+y)^2(x-y)^2\partial_x\partial_y + (x+y)^2\delta_s(\delta_s - 1) - (x-y)^2\delta_t(\delta_t - 1).$$

Miller [7] found that the parameter changing symmetry algebra of (EPD) becomes $sl_4(\mathbb{C})$. In a similar manner, we find that $\mathcal{T} \simeq sl_4(\mathbb{C})$. In fact, (H) arises from the complex wave equation

$$(\partial_t^2 - \partial_x^2 - \partial_y^2 - \partial_z^2)v = 0 \tag{4W}$$

in four-dimensional space time, by separating off two angular coordinates. Equation (EPD) also arises from this equation by another separation of variables. The symmetry algebra of (4W) is $o_6(\mathbb{C}) \simeq sl_4(\mathbb{C})$ (Miller [6]), which leads to the parameter changing symmetry algebras of (EPD) and (H). The algebra $\mathcal{T}$ will reveal a relationship between the Riemann function of (H) and Appell's F_4. In this paper, however, we shall not investigate this problem further.

3. Riemann function in a special case. Denote the Riemann function of (SEPD) (or equally of (EPD) with $a = b$) by $\hat{R}(X, Y; \Xi, \mathrm{H})$; then $R(x, y; \xi, \eta) = \hat{R}(x^2, y^2; \xi^2, \eta^2)$. Although $\hat{R}$ is already known ([2], [4]), we now give another simple calculation. We have a fibration

$$\begin{array}{ccc} \text{Möb} & \longrightarrow & \mathbb{C}^4 \\ & & \Big\downarrow \pi\text{: cross-ratio}, \qquad \dim \text{Möb} = 3, \\ & & \mathbb{C}^1 \end{array}$$

where the action of $A \in$ Möb on $\mathbb{C}^4$ is defined by $(X, Y, \Xi, \mathrm{H}) \mapsto (AX, AY, A\Xi, A\mathrm{H})$ and π is a cross-ratio

$$\pi(X, Y, \Xi, \mathrm{H}) = (X - \Xi)(Y - \mathrm{H})/(X - Y)(\Xi - \mathrm{H}).$$

On the other hand, (2), (3), and (4) show that

$$\hat{R}(X, Y; \Xi, \mathrm{H}) = \hat{R}(AX, AY; A\Xi, A\mathrm{H}), \qquad A \in \ddot{\text{Mob}},$$

so that the fibration implies that $\hat{R}$ is only a function of $s = \pi(X, Y, \Xi, \mathrm{H})$. Rewriting (SEPD) as an ordinary differential equation of an independent variable s, we find that $\hat{R}$ satisfies

$$\{s(1-s)\partial_s^2 + (\gamma - (1 + \alpha + \beta)s)\partial_s - \alpha\beta\}\hat{R} = 0, \tag{HG}$$

with $\alpha = a$, $\beta = 1 - a$, $\gamma = 1$. With this specialization of parameters, (HG) is transformed to the Legendre equation of an independent variable $t = 2s + 1$. It is interesting that (SEPD) is related to the Bessel equation on one hand and to the Legendre equation on the other hand. The condition (3) is now restated that $s = 0$ implies $\hat{R} = 1$. Hence we have the following.

THEOREM 5. *In a special case, the Riemann function of* (H) *is given by*

$$R(x, y; \xi, \eta) = F(a, 1 - a, 1; (x^2 - \xi^2)(y^2 - \eta^2)/(x^2 - y^2)(\xi^2 - \eta^2)),$$

where $F(\alpha, \beta, \gamma; s)$ *is the Gauss hypergeometric function.*

4. Group acting on the Riemann function. We turn to a generic case. The following transformations of independent variables act on the solution space of (H):

(A) $(x, y)\mapsto(\lambda x, \lambda y)$, $\lambda\in\mathbb{C}^\times$,

(B) $(x, y)\mapsto(1/x, 1/y)$,

(C) $(x, y)\mapsto(y, x)$.

(A) is a transformation corresponding to an infinitesimal one $\delta_x+\delta_y$.

The Riemann function R is invariant under the transformations of the independent variables (x, y, ξ, η) of the form

(a) $(x, y, \xi, \eta)\mapsto(\lambda x, \lambda y, \lambda\xi, \lambda\eta)$, $\lambda\in\mathbb{C}^\times$,

(b) $(x, y, \xi, \eta)\mapsto(1/x, 1/y, 1/\xi, 1/\eta)$,

(c) $(x, y, \xi, \eta)\mapsto(y, x, \eta, \xi)$,

(d) $(x, y, \xi, \eta)\mapsto(\xi, \eta, x, y)$.

Among these, (a), (b), and (c) are obtained by applying (A), (B), and (C) to (2)-(3) respectively, and (d) follows from formal self-adjointness of the operator L. By putting

$$X=x/\eta,\quad Y=y/\eta,\quad Z=\xi/\eta,$$

we can regard R as a function of X, Y, and Z because of its invariance by (a). Thus the transformations of (X, Y, Z) induced by (b), (c), and (d) are given, respectively, by

$$(X, Y, Z)\mapsto(1/X, 1/Y, 1/Z),$$

$$(X, Y, Z)\mapsto(Y/Z, X/Z, 1/Z),$$

$$(X, Y, Z)\mapsto(Z/Y, 1/Y, X/Y).$$

The group G generated by these transformations has more symmetrical generators

$$X^*:\quad (X, Y, Z)\mapsto(X, X/Z, X/Y),$$

$$Y^*:\quad (X, Y, Z)\mapsto(Y/Z, Y, Y/X),$$

$$Z^*:\quad (X, Y, Z)\mapsto(Z/Y, Z/X, Z).$$

As is easily seen, X^*, Y^*, and Z^* are involutions and commutative with each other. Hence the group $G=\langle X^*, Y^*, Z^*\rangle$ is

$$G\simeq\mathbb{Z}_2\times\mathbb{Z}_2\times\mathbb{Z}_2,\qquad |G|=8.$$

PROPOSITION 6. *The Riemann function R is a function of (X, Y, Z) and invariant under the action of G. Equations* (2)-(3) *are rewritten as*

$$L_{(X,Y)}R=0 \tag{5}$$

$$(X-Z)(Y-1)=0\quad \textit{implies } R=1. \tag{6}$$

5. Extension of a field. Consider an extension of a field $\mathbf{K}/\mathbf{k}$

$$\mathbf{K}=\mathbb{C}(X, Y, Z),$$

$$\mathbf{k}=\{f\in\mathbf{K};\ f \text{ is } G\text{-invariant}\},$$

where the action of G on $\mathbf{K}$ is one induced by the action of G on the independent variables. Then the degree of extension is given by

$$[\mathbf{K}:\mathbf{k}]=|G|=8. \tag{7}$$

With this fact in mind, we seek generators of $\mathbf{k}$ over $\mathbb{C}$. For a systematization, we denote

$$X_1=X,\quad X_2=Y,\quad X_3=Z.$$

Consider the following expressions:

$$\begin{aligned} p_i &= X_i + 1/X_i + X_k/X_j + X_j/X_k, \\ q_i &= X_i/(X_jX_k) + (X_jX_k)/X_i + 2, \end{aligned} \qquad (i = 1, 2, 3) \tag{8}$$

where (i, j, k) runs over all permutations of $(1, 2, 3)$. As is easily seen,

$$p_i, q_i \in \mathbf{k} \qquad (i = 1, 2, 3). \tag{9}$$

THEOREM 7. *For any permutation* (i, j, k) *of* $(1, 2, 3)$,

$$\mathbf{k} = \mathbb{C}(p_i, p_j, q_k).$$

Proof. Denote $\mathbf{k}' = \mathbb{C}(p_i, p_j, q_k)$. Then (9) implies $\mathbf{k}' \subset \mathbf{k}$. Hence, by (7), it suffices to show that $[\mathbf{K}:\mathbf{k}'] \leqq 8$. Putting $\theta = X_k/(X_iX_j)$, we obtain from (8),

$$\theta^2 + (2 - q_k)\theta + 1 = 0,$$

$$(1+\theta)\theta X_\nu^2 - (\theta p_\nu)X_\nu + (\theta + 1) = 0 \qquad (\nu = i, j).$$

Therefore θ, X_i and X_j are at most of degree two over $\mathbf{k}'$, $\mathbf{k}'(\theta)$ and $\mathbf{k}'(\theta, X_i)$, respectively. Hence $\mathbf{k}'(\theta, X_i, X_j) = \mathbf{k}'(X_i, X_j, X_k) = \mathbf{K}$ is at most of degree eight over $\mathbf{k}'$, namely, $[\mathbf{K}:\mathbf{k}'] \leqq 8$. □

In addition to p_i and q_i $(i = 1, 2, 3)$ defined in (8), we consider the following elements of $\mathbf{K}$:

$$\begin{aligned} r_{ij} &= X_i - 1/X_i + X_k/X_j - X_j/X_k, \\ s_i &= X_i + 1/X_i - X_k/X_j - X_j/X_k, \\ t_i &= -X_i/(X_jX_k) + (X_jX_k)/X_i. \end{aligned} \tag{10}$$

Again (i, j, k) runs over all permutations of $(1, 2, 3)$. We define an unordered triple $\langle \nu_1 1, \nu_2 2, \nu_3 3\rangle$ $(\nu_j = \pm)$ by a linear subspace of $\mathbf{K}$ over $\mathbf{k}$ of the form

$$\langle \nu_1 1, \nu_2 2, \nu_3 3\rangle = \{f \in \mathbf{K};\ X_j^*(f) = \nu_j f,\ j = 1, 2, 3\}.$$

LEMMA 8. *Let* (i, j, k) *be a permutation of* $(1, 2, 3)$. *Then*

(i) $p_i, q_i \in \mathbf{k} = \langle 1, 2, 3\rangle$, $\quad r_{ij} \in \langle i, -j, k\rangle$
$s_i \in \langle i, -j, -k\rangle$, $\quad t_i \in \langle -i, j, k\rangle$.

(ii) $p_ip_j = p_kq_k$, $\quad r_{ij}t_j = p_i(q_j - 4)$,
$r_{ij}r_{jk} = p_ip_j - 4p_k$, $\quad r_{ij}^2 = p_i^2 - 4q_k$, $\quad r_{ij}r_{ji} = q_ks_k$,
$t_i^2 = q_i(q_i - 4)$, $\quad s_i^2 = p_i^2 - 4(q_j + q_k) + 16$.

(iii) $2\delta_{X_i} = (r_{ij} + r_{ji})\partial_{p_i} + (r_{jk} - r_{ji})\partial_{p_j} + 2t_k\partial_{q_k}$,
$\delta_{X_j}(r_{ij} + r_{ji}) = \delta_{X_j}(r_{jk} - r_{ji}) = 0$, $\quad \delta_{X_j}t_k = q_k - 2$.

6. Reduction to Appell's F_4. We introduce new variables

$$r = p_1/q_3, \quad s = p_2/q_3, \quad t = q_3.$$

Since $\mathbf{k} = \mathbb{C}(p_1, p_2, q_3) = \mathbb{C}(r, s, t)$, R is regarded as a function of (r, s, t). Using Lemma 8, we can rewrite (5) as follows:

$$s_3M_1R + rstM_1R + 2M_2R + 2M_3R = 0, \tag{11}$$

where M_1, M_2, and M_3 are differential operators defined by

$$M_1 = L_{(r,s)} = \partial_r\partial_s + a(a-1)/(r-s)^2 - b(b-1)/(r+s)^2,$$

$$\begin{aligned} M_2 = {} & (1-r^2)\partial_r^2 + (1-s^2)\partial_s^2 - 2rs\partial_r\partial_s - 2r\partial_r - 2s\partial_s \\ & - 2a(a-1)/(r-s)^2 - 2b(b-1)^2/(r+s)^2, \end{aligned}$$

$$M_3 = t\{(t-4)\partial_t(\delta_t - \delta_r - \delta_s - 1) + 2\partial_t\}.$$

Since M_1R and $rstM_1R+2M_2R+2M_3R$ are G-invariant and s_3 is not an element of $\mathbf{k}$, (11) splits into two parts:

$$M_1R = M_2R + M_3R = 0. \tag{12}$$

Condition (6) is now rewritten as

$$s = 1 \text{ implies } R = 1. \tag{13}$$

In view of the explicit forms of M_j $(j = 1, 2, 3)$, if $f(r, s, t)$ is a solution of (12)-(13) (suppose f makes sense at $t = 0$), then $f(r, s, 0)$ is also a solution of (12)-(13). Hence it is reasonable to expect that R is a function depending only on (r, s), and to consider a system of partial differential equations

$$M_1u = M_2u = 0, \tag{14}$$

to which (12) reduces, if our expectation is correct. Note that if (14) has a solution satisfying the condition

$$s = 1 \text{ implies } u = 1 \quad \text{(see (13))}, \tag{13'}$$

then our expectation is actually correct, and R is given by this solution. This will be true in the rest of this section and in the next section.

We find that the substitutions

$$\begin{aligned} p &= \{(r-s)/2\}^2 = \{(x-y)(\xi-\eta)/2(xy+\xi\eta)\}^2, \\ q &= \{(r+s)/2\}^2 = \{(x+y)(\xi+\eta)/2(xy+\xi\eta)\}^2, \\ u &= p^{a/2}q^{b/2}v \end{aligned} \tag{15}$$

transform (14) to a system of partial differential equations associated with Appell's generalized hypergeometric function $F_4(\alpha, \beta, \gamma, \gamma'; p, q)$ with a specialization of parameters α, β, γ, and γ'.

Recall (see Appell and Kampé de Fériet [1], Kimura [5]) that the function F_4 is defined by a double power series

$$F_4(\alpha, \beta, \gamma, \gamma'; p, q) = \sum_{m,n=0}^{\infty} \frac{(\alpha, m+n)(\beta, m+n)}{(\gamma, m)(\gamma', n)(1, m)(1, n)} p^m q^n,$$

where (a, k) denotes a factorial function $(a, k) = a(a+1)\cdots(a+k-1)$. This power series converges in a domain

$$(p, q) \in \mathbb{C}^2, \qquad |p|^{1/2} + |q|^{1/2} < 1.$$

The function F_4 satisfies a system of partial differential equations

$$\begin{aligned} N_1v &= \{\delta_p(\delta_p+\gamma-1) - p(\delta_p+\delta_q+\alpha)(\delta_p+\delta_q+\beta)\}v = 0, \\ N_2v &= \{\delta_q(\delta_q+\gamma'-1) - q(\delta_p+\delta_q+\alpha)(\delta_p+\delta_q+\beta)\}v = 0. \end{aligned} \tag{F_4}$$

The system (F_4) can be rewritten as an integrable Pfaffian equation of rank four, so that the family of all solutions of (F_4) forms a four-dimensional vector space. Solutions of (F_4) are multivalued holomorphic functions in $\mathbb{C}^2 \backslash S$, where

$$S = \{(p, q) \in \mathbb{C}^2;\ pqD(p, q) = 0\},$$

and $D(p, q)$ is a polynomial defined by

$$D(p, q) = (p-q+1)^2 - 4p = (q-p+1)^2 - 4q. \tag{16}$$

LEMMA 9. *The substitution* (15) *transforms* (14) *to* (F_4) *with the following specialization of parameters*:

$$\alpha=(a+b)/2,\qquad \beta=(a+b+1)/2,\qquad \gamma=a+\tfrac{1}{2},\qquad \gamma'=b+\tfrac{1}{2}. \tag{17}$$

If the independent variables are real, then the condition (13′) *becomes*

$$(p, q)\in\Gamma \text{ implies } v=p^{-a/2}q^{-b/2}, \tag{18}$$

where Γ *is a part of a parabola*:

$$\Gamma=\{(p, q)\in\mathbb{R}^2;\ 0<p, q<1,\ D(p, q)=0\}.$$

We note that (14) is equivalent to the pull-back of (F_4) by a covering map

$$\mathbb{C}^2=\{(r, s)\}\to\mathbb{C}^2=\{(p, q)\},\qquad (r, s)\mapsto((r-s)^2/4,\ (r+s)^2/4),$$

which removes an apparent singularity $p+q=1$ of the Pfaffian equation associated with (F_4).

7. Integral representation. Takano [10] computed the monodoromy group of the system (F_4) on the basis of an integral representation.

THEOREM 10 (Takano [10]). *Suppose* $f(s)$ *is a solution of Gauss' hypergeometric equation*

$$\{s(1-s)\partial_s^2+[(\gamma+\gamma'-1)-(\alpha+\beta+1)s]\partial_s-\alpha\beta\}f=0, \tag{HG′}$$

and C *is a curve satisfying certain conditions*; *then an integral*

$$v(p, q)=(2\pi i)^{-1}\int_C t^{-\gamma}(1-t)^{-\gamma'}f(p/t+q/(1-t))\,dt \tag{19}$$

gives a solution of (F_4).

Hereafter we assume (17) throughout the rest of this paper. In this section, we shall observe that, under an appropriate choice of a function $f(s)$ and a curve C, a function v defined by (19) satisfies (F_4)-(18), i.e., $u=p^{a/2}q^{b/2}v$ gives a solution of (13′)-(14).

In what follows, we assume that (p, q) moves in a real domain

$$\Omega=\{(p, q)\in\mathbb{R}^2;\ 0<p, q<1,\ D(p, q)>0\}.$$

We explain how we should take a function $f(s)$ and a curve C. Under (17), (HG′) has the following characteristic exponents at each singular point:

$$s=0:\ 0,\ 1-a-b;\qquad s=1:\ 0,\ -\tfrac{1}{2};\qquad s=\infty:\ (a+b)/2,\ (a+b+1)/2.$$

Choice of $f(s)$. Let $f(s)$ be a solution of (HG′) having an exponent $-\frac{1}{2}$ at $s=1$ and normalized so that

$$f(s)=(s-1)^{-1/2}g(s),\qquad g(s)\text{: holomorphic at } s=1,\qquad g(1)=1.$$

Explicitly, $g(s)$ is given by

$$g(s)=F((a+b)/2,\ (a+b-1)/2,\ \tfrac{1}{2};\ 1-s). \tag{20}$$

If we introduce the notation

$$s=s(t)=s(t; p, q)=p/t+q/(1-t),$$
$$z_m=z_m(p, q)=\tfrac{1}{2}\{p-q+1+(-)^{m+1}\sqrt{D(p, q)}\}\quad (m=0, 1),$$
$$z_2=z_2(p, q)=p/(p-q),$$

where the square root is chosen so that $D^{1/2}>0$, then we observe

(i) $t=0$ and $t=1$ correspond to $s=\infty$,

(ii) $t=z_0$ and $t=z_1$ correspond to $s=1$,

(iii) $t=z_2$ and $t=\infty$ correspond to $s=0$.

The three points $z_m(m=0,1,2)$ are located in a following manner:

(I) $0<z_0<z_1<1$, $z_2<0$ or $z_2>1$ or $z_2=\infty$,

(II) z_0 and z_1 degenerate into a single point as $\Omega\ni(p,q)\to\Gamma$.

Choice of C. Let C be a closed Jordan curve that encircles the points z_0 and z_1 once anticlockwisely and leaves the points 0, 1, and z_2 outside.

LEMMA 11. *Let $g(s)$ and C be as above, then a function*

$$v(p,q)=(2\pi i)^{-1}\int_C t^{-a}(1-t)^{-b}(t-z_0)^{-1/2}(t-z_1)^{-1/2}g(s(t))\,dt \tag{21}$$

is a solution of (F_4)-(18).

Proof. The expression (21) is only a rewriting of (19). Since z_0 and z_1 are branch points of degree one of the integrand of (21), the curve C is a closed curve on a Riemann surface of the integrand. Hence, C satisfies a condition needed in Theorem 10, and (21) gives a solution to (F_4). We show that (21) satisfies the condition (18). In a limiting process as $(p,q)\to\Gamma$, the branch points z_0 and z_1 degenerate into a single point in an interval (0, 1), which becomes a pole of order one of the integrand. Thus, by the residue theorem, we have

$$v(p,q)=z_0^{-a}(1-z_0)^{-b}g(s(z_0)),\qquad (p,q)\in\Gamma.$$

On the other hand, if $(p,q)\in\Gamma$, (16) and (ii) show that

$$z_0=(p-q+1)/2=p^{1/2},\qquad 1-z_0=(q-p+1)/2=q^{1/2},\qquad s(z_0)=1,$$

then we obtain $v(p,q)=p^{-a/2}q^{-b/2}$, condition (18). □

We change the path of integration C into C' in a way indicated in Fig. 1. Since z_1 is a branch point of degree one of the integrand and we have an estimate

$$|\text{the integrand of (21)}|\leqq \text{const.}\,|t-z_0|^{1/2}|t-z_1|^{1/2}$$

in a neighbourhood of $[z_0,z_1]$, we obtain the following theorem.

THEOREM 12. *Let $g(s)$ be defined by* (20); *then a function*

$$v(p,q)=\pi^{-1}\int_{z_0}^{z_1} t^{-a}(1-t)^{-b}(t-z_0)^{-1/2}(z_1-t)^{-1/2}g(s(t))\,dt \tag{22}$$

is a solution of (F_4)-(18), *where the integration is taken over the interval $z_0<t<z_1$.*

As pointed out in the previous section, this theorem ensures that $R=p^{a/2}q^{b/2}v$ gives the Riemann function of (H).

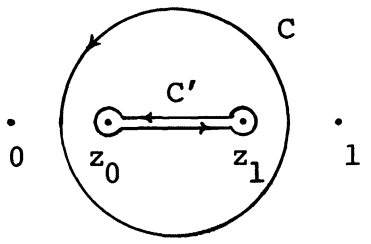


FIG. 1

8. Connection coefficients. Equation (F_4) has a fundamental set of solutions $v_j(p, q) = v_j(p, q; a, b)$ $(j = 0\text{-}3)$ in a domain $|p|^{1/2} + |q|^{1/2} < 1$ (see [1], [5]), where, in the present case,

$$
\begin{aligned}
&v_0(p, q) = F_4((a+b)/2, (a+b+1)/2, a+\tfrac{1}{2}, b+\tfrac{1}{2}; p, q),\\
&v_1(p, q) = p^{1/2-a} F_4((b-a+1)/2, 1+(b-a)/2, \tfrac{3}{2}-a, b+\tfrac{1}{2}; p, q),\\
&v_2(p, q) = q^{1/2-b} F_4((a-b+1)/2, 1+(a-b)/2, a+\tfrac{1}{2}, \tfrac{3}{2}-b; p, q),\\
&v_3(p, q) = p^{1/2-a} q^{1/2-b} F_4(1-(a+b)/2, (3-a-b)/2, \tfrac{3}{2}-a, \tfrac{3}{2}-b; p, q).
\end{aligned} \tag{23}
$$

Let $v(p, q) = v(p, q; a, b)$ be a function defined by (22). Then it can be expressed as a linear combination of v_j $(j = 0\text{-}3)$

$$
v(p, q; a, b) = \sum_{j=0}^{3} C_j(a, b) v_j(p, q; a, b). \tag{24}
$$

LEMMA 13. *The connection coefficients* C_j $(j = 0\text{-}3)$ *are meromorphic functions of* $(a, b) \in \mathbb{C}^2$ *with poles only in* $\mathcal{P}$, *where*

$$
\mathcal{P} = \{(a, b) \in \mathbb{C}^2;\ a \text{ or } b \in \tfrac{1}{2} + \mathbb{Z}\}.
$$

Proof. For a function $f(p, q)$, we define

$$
\mathbf{f} = {}^t(f^{(0)}, f^{(1)}, f^{(2)}, f^{(3)}),
$$

where

$$
f^{(0)} = f, \qquad f^{(1)} = \partial_p f, \qquad f^{(2)} = \partial_q f, \qquad f^{(3)} = \partial_p \partial_q f.
$$

For a fixed (p, q), $F_4^{(i)}(\alpha, \beta, \gamma, \gamma'; p, q)$ $(i = 0\text{-}3)$ are meromorphic functions of $(\alpha, \beta, \gamma, \gamma')$ whose poles are points for which γ or γ' is a nonpositive integer; hence (23) shows that $v_j^{(i)}$ $(i, j = 0\text{-}3)$ are meromorphic functions with poles only in $\mathcal{P}$. Since (F_4) for a function f is rewritten as an integrable Pfaffian equation for a vector $\mathbf{f}$ (see [1], [5]), a Wronskian matrix $W = (\mathbf{v}_0, \cdots, \mathbf{v}_3)$ is a nonsingular meromorphic function of (a, b) with poles only in $\mathcal{P}$. On the other hand, for a fixed (p, q), the vector $\mathbf{v}$ is an entire function of (a, b), since the Riemann function of (H) satisfies a Volterra integral equation depending entirely on the parameters (a, b). Hence the relation ${}^t(C_0, \cdots, C_3) = W^{-1}\mathbf{v}$, which follows from (24), establishes the lemma. □

In the rest of this paper, we determine the coefficients C_j. In view of Lemma 13, it suffices to determine these coefficients in some subdomains of $\mathbb{C}^2 = \{(a, b)\}$, and then continue them to whole domain by analyticity. It follows from (23) and (24) that

(1°) if $\operatorname{Re}(a) + \operatorname{Re}(b) \leqq 1$, $\operatorname{Re}(a) > \frac{1}{2}$, then

$$
v(p, p) = p^{1/2-a}\{C_1(a, b) + o(1)\} \quad \text{as } p \to 0,
$$

(2°) if $\operatorname{Re}(a) + \operatorname{Re}(b) \leqq 1$, $\operatorname{Re}(b) > \frac{1}{2}$, then

$$
v(p, p) = q^{1/2-b}\{C_2(a, b) + o(1)\} \quad \text{as } p \to 0,
$$

(3°) if $\operatorname{Re}(a) < \frac{1}{2}$, $\operatorname{Re}(b) < \frac{1}{2}$, then

$$
v(p, p) = C_0(a, b) + o(1) \quad \text{as } p \to 0,
$$

(4°) if $\operatorname{Re}(a) > \frac{1}{2}$, $\operatorname{Re}(b) > \frac{1}{2}$, then

$$
v(p, p) = p^{1-a-b}\{C_3(a, b) + o(1)\} \quad \text{as } p \to 0.
$$

This observation tells us that we have to investigate behaviours of $v(p, p)$ for small values of p in the respective cases where (a, b) lies in the four subdomains of $\mathbb{C}^2$ indicated above.

In the rest of this paper, we assume that (p, q) lies in a set

$$\Delta = \{(p, q) \in \mathbb{R}^2;\ 0 < p = q < \tfrac{3}{16}\} \subset \Omega.$$

Then $s(t)$ and z_m $(m = 0, 1, 2)$, defined before, take the forms

$$s = s(t) = s(t; p) = p/t(1-t),$$
$$z_m = z_m(p) = \tfrac{1}{2}\{1 + (-)^{m+1}\sqrt{1-4p}\} \quad (m = 0, 1),$$
$$z_2 = \infty.$$

From these, we observe that

$$\begin{aligned} &s(t) = s(1-t), \qquad 4p \leqq s(t) < 1, \qquad (z_0 < t < z_1), \\ &s(t): \quad \text{decreasing in } z_0 < t \leqq \tfrac{1}{2}, \text{ increasing in } \tfrac{1}{2} \leqq t < z_1. \end{aligned} \tag{25}$$

For $0 < p \leqq \frac{3}{16}$, z_0 and z_1 are located in the following manner:

(I′) $0 < z_0 < \frac{1}{4}$, $\frac{3}{4} < z_1 < 1$, $z_0 + z_1 = 1$, $z_0 z_1 = p$,

(II′) $z_0/p \to 1$, $(1 - z_1)/p \to 1$ as $p \to 0$.

Let (HG″) be the hypergeometric equation satisfied by $g(s)$. Then the characteristic exponents of (HG″) at the singular points $s = 0$ and 1 are

$$0 \text{ and } 1 - a - b \text{ at } s = 0; \qquad 0 \text{ and } \tfrac{1}{2} \text{ at } s = 1. \tag{26}$$

9. Determination of C_1 and C_2. An integral

$$E(a, b) = \pi^{-1} \int_0^1 s^{a-3/2}(1-s)^{-1/2} g(s)\, ds$$

is a holomorphic function of a and b in $\mathrm{Re}\,(a) + \mathrm{Re}\,(b) < 0$, $\mathrm{Re}\,(a) > \frac{1}{2}$, since (26) shows that this integral is absolutely convergent for those values of a and b.

PROPOSITION 14.

$$C_1(a, b) = E(a, b) \text{ for } \mathrm{Re}\,(a) + \mathrm{Re}\,(b) \leqq 1, \qquad \mathrm{Re}\,(a) > \tfrac{1}{2},$$
$$C_2(a, b) = E(b, a) \text{ for } \mathrm{Re}\,(a) + \mathrm{Re}\,(b) \leqq 1, \qquad \mathrm{Re}\,(b) > \tfrac{1}{2}.$$

Proof. We prove the first formula of the proposition. We assume $\mathrm{Re}\,(a) + \mathrm{Re}\,(b) \leqq 1$ and $\mathrm{Re}\,(a) > \frac{1}{2}$. Then it follows that $\mathrm{Re}\,(b) < \frac{1}{2}$. Since $\mathrm{Re}\,(1 - a - b) \geqq 0$, (26) shows that

$$|g(s)| \leqq \text{constant} \quad (0 \leqq s \leqq 1). \tag{27}$$

We divide the integral (22) into two parts whose intervals of integration are $z_0 < t < \frac{1}{2}$ and $\frac{1}{2} < t < z_1$. We denote them by $I_0(p)$ and $I_1(p)$, respectively. Namely,

$$v(p, p) = I_0(p) + I_1(p),$$

where, for $\nu = 0$ and 1,

$$I_\nu(p) = (-)^m \pi^{-1} \int_{z_\nu}^{1/2} t^{-a}(1-t)^{-b}(t - z_0)^{-1/2}(z_1 - t)^{-1/2} g(s(t))\, dt.$$

We first investigate $I_0(p)$. Substituting $t = z_0/u$, we obtain

$$I_0(p) = z_0^{1/2-a} I_2(p),$$

where $I_2(p)$ is given by

$$I_2(p) = \pi^{-1} \int_0^1 \chi(u; z_0) u^{a-3/2}(1-u)^{-1/2}(1 - z_0/u)^{-b}(z_1 - z_0/u)^{-1/2} g(s(z_0/u))\, du,$$

and $\chi(u; z_0)$ is a function having value one in $2z_0 < u < 1$ and zero outside. From (I') and (27), we have an estimate

$$|\text{the integrand of } I_2| \leqq \text{const } u^{\mathrm{Re}(a)-3/2}(1-u)^{-1/2} \quad (0 \leqq u \leqq 1),$$

the right-hand side being integrable in $0 \leqq u \leqq 1$. Formula (II') shows that

$$\text{the integrand} \to u^{a-3/2}(1-u)^{-1/2}g(u) \quad \text{as } p \to 0.$$

Hence, by Lebesgue's convergence theorem, we find that $I_2(a, b) \to E(a, b)$ as $p \to 0$. Therefore, taking the fact $z_0/p \to 1$ into account, we obtain

$$I_0(p) = p^{1/2-a}\{E(a, b) + o(1)\} \quad \text{as } p \to 0.$$

Next we consider the integral $I_1(p)$. From (27), we obtain

$$\begin{aligned} |I_1(p)| &\leqq \text{const} \int_{1/2}^{z_1} (1-t)^{-\mathrm{Re}(b)}(z_1-t)^{-1/2}\,dt \\ &= \text{const } z_1^{1/2} \int_{1/2z_1}^{1} (1-z_1t)^{-\mathrm{Re}(b)}(1-t)^{-1/2}\,dt \\ &\leqq \text{const} \int_0^1 \{1+(1-t)^{-\mathrm{Re}(b)}\}(1-t)^{-1/2}\,dt. \end{aligned}$$

Since $\mathrm{Re}\,(b) < \frac{1}{2}$, the right-hand side is finite, so that $I_1(p)$ is bounded in $0 < p < \frac{3}{16}$. Comparing the asymptotic behaviours of $I_0(p)$ and $I_1(p)$, we obtain

$$I(p) = p^{1/2-a}\{E(a, b) + o(1)\} \quad \text{as } p \to 0.$$

We combine this formula with (1°) to establish the first formula of the proposition. The second formula can be proved similarly by exchanging the role of a and b. □

LEMMA 15. $E(a, b) = \pi^{-1}2^{a-b-1}\Gamma(a-\frac{1}{2})\Gamma(\frac{1}{2}-b)/\Gamma(a-b)$.

Proof. Suppose that $\mathrm{Re}\,(a) + \mathrm{Re}\,(b) \leqq 0$, $\mathrm{Re}\,(a) > 1$, $|p| \leqq 1$, and we consider an integral

$$E(p; a, b) = \pi^{-1} \int_0^1 t^{-1/2}(1-t)^{a-3/2}g(pt)\,dt.$$

This integral converges for those (p, a, b) mentioned above, and

$$\begin{aligned} &E(1; a, b) = E(a, b), \\ &E(0; a, b) = \pi^{-1}B(\tfrac{1}{2}, a-\tfrac{1}{2}) = \pi^{-1/2}\Gamma(a-\tfrac{1}{2})/\Gamma(a). \end{aligned}$$

We observe that $E(p; a, b)$ satisfies the hypergeometric equation

$$p(1-p)E'' + \{a-(a+b+\tfrac{1}{2})p\}E - \tfrac{1}{4}(a+b)(a+b-1)E = 0.$$

Since the characteristic exponents at $p = 0$ are 0 and $1-a$ $(\mathrm{Re}\,(1-a) < 0)$,

$$E(p; a, b) = \pi^{-1/2}\Gamma(a-\tfrac{1}{2})\Gamma(a)^{-1}F((a+b)/2, (a+b-1)/2, a; p).$$

By the connection formula in Lemma 16 below, we have

$$\begin{aligned} E(p; a, b) = \hat{E}(a, b)&F((a+b)/2, (a+b-1)/2, b+\tfrac{1}{2}; 1-p) \\ &+ (1-p)^{1/2-b} \times (\text{a holomorphic function at } p = 1), \end{aligned}$$

where $\hat{E}(a, b)$ is a constant given by the right-hand side of Lemma 15. Hence, by the assumption $\mathrm{Re}\,(\frac{1}{2}-b) > 0$, we obtain $E(a, b) = E(1; a, b) = \hat{E}(a, b)$. This proves the lemma. □

A connection formula for the Gauss hypergeometric function is well known.

LEMMA 16 (*see, e.g.*, [3], [5]).

$$F(\alpha, \beta, \gamma; p) = \frac{\Gamma(\gamma)\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\Gamma(\gamma-\beta)} F(\alpha, \beta, \alpha+\beta-\gamma+1; 1-p)$$
$$+ \frac{\Gamma(\gamma)\Gamma(\alpha+\beta-\gamma)}{\Gamma(\alpha)\Gamma(\beta)} (1-p)^{\gamma-\alpha-\beta} F(\gamma-\alpha, \gamma-\beta, \gamma-\alpha-\beta+1; 1-p).$$

10. Determination of C_0 and C_3. Recall that $g(s)$ is a solution to (HG″) holomorphic in a neighbourhood of $s=1$. In the present situation, Lemma 16 is restated as follows.

LEMMA 17. *Let $h(s; a, b)$ and $A(a, b)$ be defined by*

$$h(s; a, b) = F((a+b)/2, (a+b-1)/2, a+b; s),$$
$$A(a, b) = 2^{-(a+b)}\pi^{-1}.$$

Then $g(s)$ can be represented as follows:

$$g(s) = \pi A(a, b)h(s; a, b) + \pi A(1-a, 1-b)s^{1-a-b}h(s; 1-a, 1-b). \tag{28}$$

The function $h(s: a, b)$ is not only holomorphic at $s=0$, by definition, but also bounded near $s=1$ by (26). Hence we have

$$|h(s; a, b)| \leqq \text{constant} \quad (0 \leqq s \leqq 1). \tag{29}$$

Substituting (28) into (22), we obtain

$$v(p, p) = A(a, b)J(p; a, b) + A(1-a, 1-b)J(p; 1-a, 1-b), \tag{30}$$

where $J(p; a, b)$ is given by

$$J(p; a, b) = \int_{z_0}^{z_1} t^{-a}(1-t)^{-b}(t-z_0)^{-1/2}(z_1-t)^{-1/2}h(s(t); a, b)\, dt.$$

LEMMA 18. (i) *If* $\operatorname{Re}(a) < \frac{1}{2}$ *and* $\operatorname{Re}(b) < \frac{1}{2}$, *then*

$$J(p; a, b) \to \Gamma(\tfrac{1}{2}-a)\Gamma(\tfrac{1}{2}-b)/\Gamma(1-a-b) \quad \text{as } p \to 0.$$

(ii) *If* $\operatorname{Re}(a) > \frac{1}{2}$ *and* $\operatorname{Re}(b) > \frac{1}{2}$, *then*

$$|J(p; a, b)| \leqq \text{const}\, p^{1/2-\max(\operatorname{Re}(a), \operatorname{Re}(b))} \quad (0 < p < \tfrac{3}{16}).$$

Proof. (i) Let $t = z_1 u + z_0(1-u)$, then $J(p; a, b)$ is rewritten as

$$J(p; a, b) = \int_0^1 (uz_1 + (1-u)z_0)^{-a}(uz_0 + (1-u)z_1)^{-b}$$
$$\times u^{-1/2}(1-u)^{-1/2}h(s(uz_1 + (1-u)z_0); a, b)\, du.$$

Hence (I′) and (29) shows that the integrand satisfies

$$|\text{the integrand}| \leqq \text{const}\ (1+(u/2))^{-\operatorname{Re}(a)}(1+(u/2))^{-\operatorname{Re}(b)}u^{-1/2}(1-u)^{-1/2}.$$

Since $\operatorname{Re}(a) < \frac{1}{2}$ and $\operatorname{Re}(b) < \frac{1}{2}$, the right-hand side is integrable in $0 < u < 1$. Moreover, $z_0 \to 0$, $z_1 \to 1$ and $s(uz_1 + (1-u)z_0) \to 0$ as $p \to 0$, so that we have

$$\text{the integrand} \to u^{-a-1/2}(1-u)^{-b-1/2} \quad \text{as } p \to 0.$$

Hence, by Lebesgue's convergence theorem, we obtain

$$J(p; a, b) \to B(\tfrac{1}{2}-a, \tfrac{1}{2}-b) = \Gamma(\tfrac{1}{2}-a)\Gamma(\tfrac{1}{2}-b)/\Gamma(1-a-b),$$

where $B(\cdot, \cdot)$ is the beta function. This proves assertion (i).

(ii) If $\mathrm{Re}\,(a)$, $\mathrm{Re}\,(b)>\frac{1}{2}$, then $J(p; a, b)$ is estimated as follows:

$$\begin{aligned}|J(p; a, b)| &\leqq \text{const} \int_{z_0}^{1/2} t^{-\mathrm{Re}(a)}(t-z_0)^{-1/2}\,dt \\ &\quad + \text{const} \int_{1/2}^{z_1} (1-t)^{-\mathrm{Re}(b)}(z_1-t)^{-1/2}\,dt \\ &\leqq \text{const} \int_{z_0}^{1/2} \{t^{-\mathrm{Re}(a)}+t^{-\mathrm{Re}(b)}\}(t-z_0)^{-1/2}\,dt \\ &\leqq \text{const}\, z_0^{1/2-\max(\mathrm{Re}(a),\mathrm{Re}(b))} \int_1^{\infty} u^{-\min(\mathrm{Re}(a),\mathrm{Re}(b))}(u-1)^{-1/2}\,du \\ &\leqq \text{const}\, p^{1/2-\max(\mathrm{Re}(a),\mathrm{Re}(b))}.\end{aligned}$$

Here we note that the third inequality follows from the substitution $t=z_0u$ in the second line, and the integral in the third line is finite, since $\mathrm{Re}\,(a)$, $\mathrm{Re}\,(b)>\frac{1}{2}$. Hence the assertion (ii) is proved. □

PROPOSITION 19.

$$C_0(a, b)=\pi^{-1}2^{-(a+b)}\Gamma(\tfrac{1}{2}-a)\Gamma(\tfrac{1}{2}-b)/\Gamma(1-a-b),$$
$$C_3(a, b)=\pi^{-1}2^{a+b-2}\Gamma(a-\tfrac{1}{2})\Gamma(b-\tfrac{1}{2})/\Gamma(a+b-1).$$

Proof. We prove the first formula. Suppose $\mathrm{Re}\,(a)<1/2$ and $\mathrm{Re}\,(b)<1/2$. Applying Lemma 18 (ii) with a and b replaced by $1-a$ and $1-b$, respectively, we obtain an estimate

$$|p^{1-a-b}J(p; a, b)| \leqq \text{const}\, p^{1/2-\max(\mathrm{Re}(a),\mathrm{Re}(b))} = o(1) \quad \text{as } p\to 0.$$

Hence (30) and Lemma 18(i) shows that

$$v(p, p)=A(a, b)\Gamma(\tfrac{1}{2}-a)\Gamma(\tfrac{1}{2}-b)/\Gamma(1-a-b) \quad \text{as } p\to 0.$$

Comparing this formula with (3°), we obtain the first formula of the proposition in a domain $\mathrm{Re}\,(a)$, $\mathrm{Re}\,(b)<\frac{1}{2}$, and then in a whole domain by analyticity (see Lemma 13). Next we prove the second formula. Suppose that $\mathrm{Re}\,(a)$, $\mathrm{Re}\,(b)>\frac{1}{2}$. Applying Lemma 18(ii), we obtain

$$J(p; a, b)=o(|p^{1-a-b}|) \quad \text{as } p\to 0.$$

Hence (30) and Lemma 18(i) with a and b replaced by $1-a$ and $1-b$ shows that

$$v(p, p)=p^{1-a-b}\{A(a, b)\Gamma(a-\tfrac{1}{2})\Gamma(b-\tfrac{1}{2})/\Gamma(a+b-1)+o(1)\},$$

Comparing this formula with (4°), we obtain the second formula. □

11. Conclusion. We have proved the following theorem.

THEOREM 20. *The Riemann function of* (H) *is given by*

$$R(x, y; \xi, \eta)=\sum_{j=0}^{3} C_j(a, b)u_j(p, q; a, b),$$

where

$$p=\left\{\frac{(x-y)(\xi-\eta)}{2(xy+\xi\eta)}\right\}^2, \qquad q=\left\{\frac{(x+y)(\xi+\eta)}{2(xy+\xi\nu)}\right\}^2,$$

$$u_0=u(p, q; a, b), \qquad C_0=C(a, b),$$
$$u_1=u(p, q; 1-a, b), \qquad C_1=C(1-a, b),$$
$$u_2=u(p, q; a, 1-b), \qquad C_2=C(a, 1-b),$$
$$u_3=u(p, q; 1-a, 1-b), \qquad C_3=C(1-a, 1-b),$$

and

$$u(p, q; a, b) = p^{a/2}q^{b/2}F_4((a+b)/2, (a+b+1)/2, a+\tfrac{1}{2}, b+\tfrac{1}{2}; p, q),$$
$$C(a, b) = \pi^{-1}2^{-a-b}\Gamma(\tfrac{1}{2}-a)\Gamma(\tfrac{1}{2}-b)/\Gamma(1-a-b).$$

Remark 21. We look back upon this note and give some comments.

(i) Arguments in § 4 and in a part of § 6 are rather formal and heuristic, but they are justified later. See a comment following Theorem 12.

(ii) Theorem 20 shows that the Riemann function of (H) is represented in a very symmetric manner. But the determination of the connection coefficients C_1 and C_2 (§ 9) and that of C_0 and C_3 (§ 10) are not so symmetric. It is because a reduction to (F_4)-(18) and an integral representation of its solution break down a symmetry.

12. Appendix. We give a proof of Lemma 2. Such a statement was not in print, as far as we are aware.

The most important property of the Lie algebra $\mathcal{W}$ is as follows:

(*) *If $f, g \in \mathcal{W}$ and $[f, g] = 0$, then f and g are linearly dependent.*

Let $\mathfrak{g} \neq \{0\}$ be a finite-dimensional subalgebra of $\mathcal{W}$. Then the following two cases occur.

Case 1. For every $f \in \mathfrak{g}$, $ad(f) = [f, \cdot] \in \mathrm{End}(\mathfrak{g})$ is nilpotent.

Case 2. For some $f \in \mathfrak{g}$, $ad(f)$ is not nilpotent.

In Case 2, by considering a constant multiple of f, if necessary, we may assume that $ad(f)$ has an eigenvalue 1. Let e be an eigenvector of $ad(f)$ corresponding to the eigenvalue 1. Note that the following commutation relation holds:

(**) $$ad(e)ad(f) = (ad(f) - 1)ad(e).$$

Case 1. $\dim \mathfrak{g} = 1$.

Proof. For any $f \in \mathfrak{g}$, there exists an $m \geqq 1$ such that $ad(f)^m g = [f, ad(f)^{m-1}g] = 0$ for every $g \in \mathfrak{g}$. Hence, (*) implies that $ad(f)^{m-1}g = cf$ for some constant c. If $m \geqq 2$, then this shows that $ad(f')f = cf$, where $f' = -ad(f)^{m-2}g$. By the nilpotency of $ad(f')$, we have $c = 0$. Hence, $ad(f)^{m-1}g = 0$ for every $g \in \mathfrak{g}$. By repeating this argument, we have $ad(f)g = 0$ $(g \in \mathfrak{g})$. Hence (*) implies that $\mathfrak{g} = \mathbb{C}f$, which proves the assertion. □

Case 2. Let f and $e \in \mathfrak{g}$ be as above.

CLAIM 1. *Eigenvalues of $ad(f) \in \mathrm{End}(\mathfrak{g})$ are at most 0 and ± 1.*

Proof. Let λ be an eigenvalue other than 0 and 1 (if it exists). If $m + \lambda \neq 1$ held for every $m \in \mathbb{N}$, then infinitely many numbers $m + \lambda$ $(m \in \mathbb{N})$ would become eigenvalues of $ad(f)$. (Indeed, this assertion is apparently true for $m = 0$. If it is true for $m = n$, and let e_n be an eigenvector corresponding to $n + \lambda$ $(\neq 1)$, then e and e_n are linearly independent. Here we recall that e is an eigenvector of $ad(f)$ corresponding to 1. Hence (*) implies that $e_{n+1} := [e, e_n] \neq 0$ and $ad(f)e_{n+1} = (n + 1 + \lambda)e_{n+1}$, which shows that the assertion is also true for $m = n + 1$.) This contradicts the finite-dimensionality of $\mathfrak{g}$. Hence, there exists an $m \in \mathbb{N}$ such that $m + \lambda = 1$. Similarly by exchanging the role of 1 and λ, we see that there exists an $n \in \mathbb{N}$ such that $1 + n\lambda = \lambda$. Thus we have $\lambda = 1 - m = (1 - n)^{-1}$, which can happen only when $m = n = 2$ and $\lambda = -1$. This establishes the claim. □

CLAIM 2. *The generalized eigenspace of $ad(f)$ corresponding to a nonzero eigenvalue is one-dimensional.*

Proof. We show this claim for the eigenvalue 1. Similarly, we can prove it for -1, if it is an eigenvalue. If g is any generalized eigenvector corresponding to 1, then there exists an $m \in \mathbb{N}$ such that $(ad(f) - 1)^m g = 0$. Operating $ad(e)$ on this equality and using

(**), we have $(ad(f)-2)^m[e, g]=0$. By Claim 1, 2 is not an eigenvalue of $ad(f)$, so that $[e, g]=0$, which, combined with (*), shows that $g \in \mathbb{C}e$. This establishes the claim. □

CLAIM 3. *The generalized eigenspace of $ad(f)$ corresponding to the eigenvalue 0 is also one-dimensional.*

Proof. Let g be any generalized eigenvector corresponding to 0. Note that $[e, g] \neq 0$. There exists an $m \in \mathbb{N}$ such that $ad(f)^m g=0$. Operating $ad(e)$ on this equality and using (**), we have $(ad(f)-1)^m[e, g]=0$. This shows that $[e, g]$ is a generalized eigenvector of $ad(f)$ corresponding to the eigenvalue 1. Hence, by Claim 2, there exists a constant c such that $[g, e]=ce$. Taking $ad(f)e=e$ into account, we obtain $[g-cf, e]=0$. Hence, by (*), $g-cf=c'e$ for some constant c'. However, $g-cf$ is a generalized eigenvector corresponding to 0, and $c'e$ is a one corresponding to 1, so that $c'=0$ and $g=cf$. This establishes the claim. □

By using these claims, we can easily show that, in Case 2, the subalgebra $\mathfrak{g}$ must be noncommutative Lie algebra of dimension 2 or 3. Conversely, the examples $\mathfrak{g}=\mathbb{C} \oplus \mathbb{C}x$ and $\mathfrak{g}=\mathbb{C} \oplus \mathbb{C}x \oplus \mathbb{C}x^2$ ensure the actual existence of such subalgebras, where $\oplus$ denotes the direct sum as a linear space. This proof is due to a discussion with M. Furuta.

Acknowledgment. The author wishes to thank Professor W. Miller, Jr. for several helpful suggestions about the symmetries of the differential equations examined in this paper.

REFERENCES

[1] P. APPELL AND J. KAMPÉ DE FÉRIET, *Fonctions hypergéométriques et hypersphériques—Polynomes d'Hermite*, Gauthier-Villars, Paris, 1926.

[2] G. DARBOUX, *Leçons sur la théorie générale des surfaces*, Gauthier-Villars, Paris, 1915.

[3] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER, AND F. G. TRICOMI, *Higher Transcendental Functions, Vol.* I, McGraw-Hill, New York, 1953.

[4] IWANAMI SÛGAKU ZITEN, *Encyclopedic Dictionary of Mathematics*, 3*rd ed.*, Tokyo, 1985, English transl., MIT Press, Cambridge, MA, 1977.

[5] T. KIMURA, *Hypergeometric functions of two variables*, Seminar Notes in Mathematics, Univ. of Tokyo, 1973.

[6] W. MILLER, JR., *Symmetry and separation of variables*, Encyclopedia of Mathematics and Its Application, Vol. 4, Addison-Wesley, Reading, MA, 1977.

[7] ———, *Symmetries of differential equations, the hypergeometric and Euler-Darboux equation*, SIAM J. Math. Anal., 4 (1973), pp. 314-328.

[8] E. G. KALNINS AND W. MILLER, JR., *Lie theory and the wave equation in space-time* 1-5, J. Math. Phys., 18 (1977), pp. 1-16, 271-280; 19 (1978), pp. 1233-1246, 1247-1257; SIAM J. Math. Anal., 9 (1978), pp. 12-32.

[9] K. OKAMOTO, *Echelles et l'Equations de Toda*, preprint.

[10] K. TAKANO, *Monodromy group of the system for Appel's F_4*, Funkcialaj Ekvacioj, 23 (1980), pp. 97-122.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

011

UNIFORMLY VALID COMPOSITE EXPANSIONS FOR LAPLACE INTEGRALS*

LINDSAY A. SKINNER†

Abstract. A uniformly valid asymptotic expansion for integrals of the form

$$F(t,\nu)=e^{\nu h(t)}\int_t^\infty e^{-\nu h(x)}g(x)x^{\alpha-1}\,dx,$$

where $0<\alpha\leqq 1$ and $h(x)$ has a zero of order $r\geqq 1$ at $x=0$, is established. The result, which generalizes a well-known one for $h(x)=x$, also confirms the formal matched asymptotic expansion solution of the equivalent singular perturbation problem $-y'+\nu h'(t)y=g(t)t^{\alpha-1}$, $y(\infty)=0$. Comparable matched expansion results are derived for related integrals, including one from Bessel function theory, which do not satisfy differential equations.

Key words. matched asymptotic expansions, coalescing critical points

AMS(MOS) subject classification. 41A60

1. Introduction. The first part of this paper is concerned with the asymptotic evaluation of integrals of the form

$$F(t,\nu)=e^{\nu h(t)}\int_t^\infty e^{-\nu h(x)}g(x)x^{\alpha-1}\,dx. \tag{1.1}$$

It is assumed that $0<\alpha\leqq 1$ and $g(x)$, $h(x)\in C^\infty[0,\infty)$. Also, $h'(x)>0$ for $x>0$, but $h(x)$ has a zero of order $r\geqq 1$ at $x=0$. In addition, $g^{(n)}(x)$, $k^{(n)}(x)=O(1)$ as $x\to\infty$, where $k(x)=1/h'(x)$. Thus, in the integral, $h(x)$ has its minimum at the endpoint $x=t$, and this point coalesces with the singular point $x=0$ as $t\to 0^+$. Also, $h'(x)=x^{r-1}a(x)$ where $a(x)>0$ for $x\geqq 0$.

Our initial objective is to establish an asymptotic expansion for $F(t,\nu)$ that is uniformly valid for $0\leqq t<\infty$ as $\nu\to\infty$. The result is a generalization of a well-known one [1], [4], [8], [9] for $h(x)=x$. Following this we take up variations of (1.1) in which the integrand may depend on t. As an example, for

$$A(t,\nu)=e^{\nu(\tanh t-t)}\int_0^\infty e^{-\nu(\sinh x\,\operatorname{sech} t-x)}x^{\alpha-1}\,dx, \tag{1.2}$$

which is related to the Anger function [5], we establish

$$\begin{aligned}A(t,\nu)=\nu^{-\alpha/3}Z_0(\nu^{1/3}t)+(2\pi\nu)^{-1/2}t^{\alpha-1}[(\tanh t)^{-1/2}-t^{-1/2}]\\ +\tfrac{1}{6}\nu^{-(2+\alpha)/3}V(\nu^{1/3}t)+O(\nu^{-(4+\alpha)/3})\end{aligned} \tag{1.3}$$

uniformly for $0\leqq t<\infty$. In this formula

$$V(T)=\tfrac{1}{4}TZ_4(T)-T^3Z_2(T)-(2\pi)^{1/2}T^{\alpha+1/2} \tag{1.4}$$

and

$$Z_k(T)=\int_{-T}^\infty X^k e^{-(X^3+3TX^2)/6}(X+T)^{\alpha-1}\,dX. \tag{1.5}$$

* Received by the editors November 3, 1986; accepted for publication (in revised form) July 21, 1987.
† Department of Mathematical Sciences, University of Wisconsin, Milwaukee, Wisconsin 53201.

Like (1.1), the integral for $A(t, \nu)$ involves coalescing critical points. There are saddle points at $x = \pm t$, and the algebraic singularity at $x = 0$. A uniform approximation asymptotically equivalent to (1.3) could be obtained by the method of Chester, Friedman, and Ursell [3] and Bleistein [2]. The result would not be so simple, however, nor would the proof of its validity. Details, which involve the change of variable defined by $\frac{1}{3}z^3 + \zeta(t)z = \sinh x \operatorname{sech} t - x$, where $\zeta(t) = [\frac{3}{2}(t - \tanh t)]^{2/3}$, are given in [5] for the case $\alpha = 1$. A comparable expansion can be obtained for $F(t, \nu)$ by introducing $z = [h(x)]^{1/r}$ and expanding the resulting coefficient of $z^{\alpha-1}\exp(-\nu z^r)$ about $z = [h(t)]^{1/r}$.

The proof of (1.3), and analogous expansions for other integrals, including $F(t, \nu)$, is accomplished in two steps. In the first step we prove the existence of a preliminary expansion comparable to the type of expansion established for Laplace integrals in [6]. In the second step we show that Theorem 2, given in the next section, can be applied to the individual terms of the preliminary expansion, and that this yields the final result. As a bonus we learn in the end that actual computations can be done directly, without a preliminary stage, using the method of matched asymptotic expansions.

2. Basic results. To begin our analysis, it is appropriate, but not essential, to observe that $F(t, \nu)$ is the solution of the singular perturbation problem

$$-y' + \nu h'(t)y = g(t)t^{\alpha-1}, \qquad y(\infty) = 0. \tag{2.1}$$

From here the desired uniform expansion can be readily determined formally by matching inner and outer expansions. Bearing in mind that $h'(t) = t^{r-1}a(t)$, if we write the N-term outer expansion as

$$O_N F(t, \nu) = \nu^{-1/r} \sum_{n=0}^{N-1} \nu^{-n/r} y_n(t); \tag{2.2}$$

then, obviously,

$$y_{r-1}(t) = t^{\alpha-1}g(t)/h'(t), \tag{2.3}$$

$$y_{kr-1}(t) = y'_{kr-r-1}(t)/h'(t), \qquad k \geqq 2, \tag{2.4}$$

and $y_n(t) = 0$ if $n \neq kr - 1$. The corresponding formal N-term inner expansion has the form

$$I_N F(t, \nu) = \nu^{-\alpha/r} \sum_{m=0}^{N-1} \nu^{-m/r} Y_m(\nu^{1/r}t), \tag{2.5}$$

and it is easy to see

$$Y_0(T) = g(0)Q_0(T), \tag{2.6}$$

$$Y_1(T) = g'(0)Q_1(T) - \frac{1}{r+1}g(0)a'(0)[Q_{r+1}(T) - T^{r+1}Q_0(T)], \tag{2.7}$$

where

$$Q_k(T) = e^{[a(0)/r]T^r} \int_T^\infty e^{-[a(0)/r]X^r} X^{\alpha-1+k}\, dX. \tag{2.8}$$

Additional terms of (2.5) can be obtained by the method developed in § 4.

The coefficients in (2.2) have asymptotic expansions of the form

$$y_n(t) \sim t^{\alpha-n-1} \sum_{m=0}^{\infty} c_{mn} t^m, \qquad t \to 0^+. \tag{2.9}$$

Presumably, therefore,

$$Y_m(T) \sim T^{\alpha+m-1} \sum_{n=0}^{\infty} c_{mn} T^{-n}, \qquad T \to \infty, \tag{2.10}$$

so that $O_N I_N F(t, \nu)$ is the same as

$$I_N O_N F(t, \nu) = \nu^{-\alpha/r} \sum_{m=0}^{N-1} \nu^{-m/r} \sum_{n=0}^{N-1} c_{mn} (\nu^{1/r} t)^{\alpha+m-n-1}. \tag{2.11}$$

Then the composite $C_N F(t, \nu)$, where $C_N = O_N + I_N - O_N I_N$, has the same N-term inner and outer expansions as $F(t, \nu)$. Thus we come to the following theorem.

THEOREM 1. *Under the assumptions stated in the three sentences just below* (1.1), $F(t, \nu) = C_N F(t, \nu) + O(\nu^{-(N+\alpha)/r})$ *uniformly for* $0 \leqq t < \infty$, *as* $\nu \to \infty$. *In other words,*

$$F(t, \nu) = \nu^{-\alpha/r} \sum_{n=0}^{N-1} \nu^{-n/r} [q_n(\nu^{1/r} t) + \nu^{-(1-\alpha)/r} p_n(t)] + O(\nu^{-(N+\alpha)/r}), \tag{2.12}$$

where

$$p_n(t) = y_n(t) - t^{\alpha-n-1} \sum_{m=0}^{n} c_{mn} t^m, \tag{2.13}$$

$$q_m(T) = Y_m(T) - T^{\alpha+m-1} \sum_{n=0}^{m-1} c_{mn} T^{-n}, \tag{2.14}$$

for any $N \geqq 1$.

For $r = 1$, and $N = 1$, (2.12) says

$$F(t, \nu) = \lambda^{-\alpha} g(0)\, e^{-\lambda t} \Gamma(\alpha, \lambda t) + \nu^{-1} [G(t) - G(0)] + O(\nu^{-\alpha-1}), \tag{2.15}$$

where $\lambda = a(0)\nu$, $G(t) = g(t)/a(t)$ and $\Gamma(\alpha, T)$ is the complementary incomplete gamma function [5]. Note also that for any $r > 1$, $p_n(t) = 0$ and $q_n(T) = Y_n(T)$ for $n \leqq r - 1$.

To prove Theorem 1 we need a generalization of the composite expansion theory given in [6] and [7]. We will use the same notation. Thus

$$\phi^{[m,-n]}(t, T) = \frac{1}{(m!)(n!)} \left(\frac{\partial}{\partial t}\right)^m \left(-T^2 \frac{\partial}{\partial T}\right)^n \phi(t, T) \tag{2.16}$$

and $\phi(t, T) \in C^\infty([0, b] \times [1, \infty])$ means $\tilde{\phi}(t, T) \in C^\infty([0, b] \times [0, 1])$, where $\tilde{\phi}(t, T) = \phi(t, 1/T)$. Also, $f^{[n]}(x, X, t) = f^{[n,0,0]}(x, X, t)$, and $f(x, X, t) = o(X^{-\infty})$ means $f(x, X, t) = o(X^{-n})$ for any n. The corollary following Theorem 2 is essentially Corollary 1 in [6].

THEOREM 2. *Let* $f(t, T) = T^{\alpha-1} \phi(t, T)$ *where* $0 < \alpha \leqq 1$. *If* $f(t, T) \in C^\infty([0, b] \times [0, 1])$ *and* $\phi(t, T) \in C^\infty([0, b] \times [1, \infty])$, *then*

$$f(t, \mu t) = \sum_{n=0}^{N-1} \mu^{-n} [v_n(\mu t) + \mu^{\alpha-1} u_n(t)] + O(\mu^{-N}) \tag{2.17}$$

uniformly for $0 \leqq t \leqq b$, *as* $\mu \to \infty$, *where*

$$u_n(t) = t^{\alpha-n-1} \left[\phi^{[0,-n]}(t, \infty) - \sum_{m=0}^{n} \phi^{[m,-n]}(0, \infty) t^m \right], \tag{2.18}$$

$$v_m(T) = T^{\alpha+m-1} \left[\phi^{[m,0]}(0, T) - \sum_{n=0}^{m-1} \phi^{[m,-n]}(0, \infty) T^{-n} \right]. \tag{2.19}$$

Proof. From $\phi(t, T) \in C^\infty([0, b]\times[1, \infty])$, by Taylor's theorem,

$$\phi(t, T) = \sum_{n=0}^{N-1} T^{-n}\phi^{[0,-n]}(t, \infty) + O(T^{-N}) \tag{2.20}$$

uniformly for $0 \leqq t \leqq b$ as $T \to \infty$. Therefore

$$f(t, T) = T^{\alpha-1} \sum_{n=0}^{N-1} T^{-n}\phi^{[0,-n]}(t, \infty) + O(T^{-N}) \tag{2.21}$$

uniformly for $0 \leqq t \leqq b$ as $T \to \infty$. Similarly, $\phi(t, T) \in C^\infty([0, b]\times[1, \infty])$ implies

$$f(t, T) = T^{\alpha-1} \sum_{n=0}^{N-1} t^m\phi^{[m,0]}(0, T) + O(t^N) \tag{2.22}$$

uniformly for $1 \leqq T \leqq \infty$, as $t \to 0^+$. In fact, (2.22) holds uniformly for $0 \leqq T \leqq \infty$, since, in addition, $f(t, T) \in C^\infty([0, b]\times[0, 1])$. The remainder of the proof of this theorem parallels the proof of Theorem 1 in [6], which is the current theorem with $\alpha = 1$, and therefore is omitted.

COROLLARY 1. *If* $f(x, X, t) \in C^\infty([0, b]\times[0, \infty]\times[0, c])$ *and if* $f(x, X, t)$ *is uniformly* $o(X^{-\infty})$ *as* $X \to \infty$, *then*

$$f(x, \mu x, t) = \sum_{k=0}^{N-1} \mu^{-k}[(\mu x)^k f^{[k]}(0, \mu x, t)] + O(\mu^{-N}) \tag{2.23}$$

uniformly for all $(x, t) \in [0, b]\times[0, c]$ *as* $\mu \to \infty$.

3. Proof of Theorem 1. For the first step in proving Theorem 1 observe that straightforward integration by parts shows

$$F(t, \nu) = \nu^{-1/r} \sum_{n=0}^{N-1} \nu^{-n/r} y_n(t) + O(\nu^{-(N+1)/r}) \tag{3.1}$$

uniformly for $c \leqq t < \infty$, for any $c > 0$. The coefficients here are the same as in (2.2). Next,

$$e^{\nu h(t)} \int_{b+t}^{\infty} e^{-\nu h(x)} g(x) x^{\alpha-1}\, dx = e^{-\nu[h(b+t)-h(t)]} F(b+t, \nu) \tag{3.2}$$

and hence, for any $b > 0$,

$$F(t, \nu) = I(t, \nu) + o(\nu^{-\infty}) \tag{3.3}$$

uniformly for $0 \leqq t \leqq b$, where

$$I(t, \nu) = \int_0^b e^{-\nu[h(x+t)-h(t)]} g(x+t)(x+t)^{\alpha-1}\, dx. \tag{3.4}$$

Let $\psi(x, t) = x^{-r}[h(x+t) - h(t) - \kappa(x, t)]$, where

$$\kappa(x, t) = \sum_{k=1}^{r-1} x^k h^{[k]}(t). \tag{3.5}$$

Then $\psi(0, 0) = a(0)r! > 0$, so, for $b > 0$ sufficiently small, $\psi(x, t) > 0$ on $[0, b]\times[0, b]$. Therefore

$$f(x, X, t) = g(x+t) \exp[-X^r \psi(x, t)] \tag{3.6}$$

is in $C^\infty([0, b]\times[0, \infty]\times[0, b])$ and we can apply Corollary 1 to expand $f(x, \nu^{1/r}x, t)$. Upon substituting into

$$I(t, \nu) = \int_0^b f(x, \nu^{1/r}x, t)\, e^{-\nu\kappa(x,t)}(x+t)^{\alpha-1}\, dx, \tag{3.7}$$

and returning to (3.3), we find

$$F(t, \nu) = \nu^{-\alpha/r} \sum_{k=0}^{N-1} \nu^{-k/r} Q_k(t, \nu^{1/r}t) + O(\nu^{-(N+\alpha)/r}) \tag{3.8}$$

uniformly for $0 \leqq t \leqq b$, where

$$Q_k(t, T) = \int_0^\infty X^n\, e^{-\sigma(X,T,t)} q_n(X, t)(X+T)^{\alpha-1}\, dX. \tag{3.9}$$

Here we have introduced

$$\sigma(X, T, t) = \sum_{k=1}^{r} X^k T^{r-k} a_k(t) \tag{3.10}$$

where $a_k(t) = t^{k-r}h^{[k]}(t)$. Also

$$q_0(X, t) = g(t), \qquad q_1(X, t) = g'(t) - g(t)a_r(t)X^r, \tag{3.11}$$

and in general

$$q_n(X, t) = f^{[n]}(0, X, t) \exp[a_r(t)X^r] \tag{3.12}$$

is a polynomial of degree n in X^r.

Expansion (3.8) is to be expected from our work in [6]. In general it is not uniformly valid for $0 \leqq t < \infty$. Nevertheless, from (3.8) we will now derive (2.12), which is uniformly valid for $0 \leqq t < \infty$.

To get from (3.7) to (3.8) we had to have $Q_n(t, T) = O(1)$ on $[0, b]\times[0, \infty)$. To see that this holds, it suffices to check that

$$f(t, T) = \int_T^\infty X^k\, e^{-\sigma(X,T,t)}(X+T)^{\alpha-1}\, dX \tag{3.13}$$

is uniformly $O(1)$ in t as $T \to \infty$ for any $k \geqq 1$. If we substitute TX for X in (3.13), we get

$$f(t, T) = T^{k+\alpha} \int_0^\infty X^k\, e^{-T^r\rho(X,t)}(X+1)^{\alpha-1}\, dX \tag{3.14}$$

where $\rho(X, t) = \sigma(X, 1, t)$. Also, since

$$a_k(0) = \binom{r}{k} h^{[r]}(0) > 0, \tag{3.15}$$

we may presume $a_k(t) > 0$ on $[0, b]$, and thus $\rho^{[1]}(X, t) > 0$ on $[0, \infty)\times[0, b]$. Therefore, $f(t, T)$ has a uniformly valid expansion of the form

$$f(t, T) \sim T^{\alpha-1} \sum_{n=0}^{\infty} c_n(t) T^{-n}, \qquad T \to \infty, \tag{3.16}$$

and each $c_n(t) \in C^\infty[0, b]$. But this says more than just $f(t, T) = O(1)$ as $T \to \infty$. It shows that $f(t, T)$, and therefore each $Q_k(t, T)$, satisfies the hypotheses of Theorem 2. Hence,

$$Q_k(t, \mu t) = \sum_{n=0}^{N-1} \mu^{-n}[v_{kn}(\mu t) + \mu^{\alpha-1}u_{kn}(t)] + O(\mu^{-N}) \tag{3.17}$$

uniformly for $0 \leqq t \leqq b$, where $u_{kn}(t)$, $v_{kn}(T)$ are defined, for each k, by (2.18), (2.19). Upon substituting (3.17) into (3.8), we have

$$F(t, \nu) = \nu^{-\alpha/r} \sum_{k=0}^{N-1} \nu^{-k/r} [q_k(\nu^{1/r} t) + \nu^{-(1-\alpha)r} p_k(t)] + O(\nu^{-(N+\alpha)/r}) \tag{3.18}$$

for $0 \leqq t \leqq b$, where

$$q_k(T) = \sum_{n=0}^{k} v_{n,k-n}(T), \qquad p_k(t) = \sum_{n=0}^{k} u_{n,k-n}(t). \tag{3.19}$$

It remains to see, in view of (3.1), and (2.19), that (3.18) actually holds for $0 \leqq t < \infty$, and, furthermore, that the functions $p_k(t)$ and $q_k(T)$ defined above are the same as the ones defined by (2.13) and (2.14). But these are fairly routine items, and we shall omit the details. This completes the proof of Theorem 1.

4. Related integrals. It is a straightforward matter to modify the proof of Theorem 1 to cover integrals of the more general form

$$\phi(t, \nu) = e^{\nu h(t,t)} \int_t^\infty e^{-\nu h(x,t)} g(x, t) x^{\alpha-1} \, dx. \tag{4.1}$$

Assume $g(x, t)$, $h(x, t) \in C^\infty([0, \infty) \times [0, \infty))$, $h^{[1]}(x, t) = x^{r-1} a(x, t)$ with $a(x, t) > 0$, and $g^{[n]}(x, t)$, $k^{[n]}(x, t) = O(1)$ uniformly in t as $x \to \infty$, where $k(x, t) = 1/h^{[1]}(x, t)$. The same integration by parts process that led to (3.1) now yields

$$\phi(t, \nu) = \nu^{-1/r} \sum_{n=0}^{N-1} \nu^{-n/r} y_n(t, t) + O(\nu^{-(N+1)/r}) \tag{4.2}$$

for $t \geqq c > 0$, where, in analogy with (2.3) and (2.4),

$$y_{r-1}(x, t) = x^{\alpha-1} g(x, t)/h^{[1]}(x, t), \tag{4.3}$$

$$y_{kr-1}(x, t) = y^{[1]}_{kr-r-1}(x, t)/h^{[1]}(x, t), \qquad k \geqq 2 \tag{4.4}$$

and $y_n(x, t) = 0$ for $n \neq kr - 1$. Similarly, for $0 \leqq t \leqq b$ we again get (3.8), except now $a_k(t) = t^{k-r} h^{[k]}(t, t)$. Also $g^{[k]}(t, t)$ replaces $g^{[k]}(t)$ in the formulas for $q_k(X, t)$. Finally, we still have $a_k(0) > 0$, so we can still apply Theorem 2 to the (modified) terms of (3.8). The resulting expansion has the same form as (3.18), and like (3.18) it is uniformly valid for $0 \leqq t < \infty$.

Unlike (1.1), in the case of (4.1) we do not, in general, have an equivalent differential equation problem. From the above discussion, however, it is clear that, nevertheless, $\phi(t, \nu)$ does have matching inner and outer expansions. Furthermore, we can calculate these expansions directly. Indeed, the outer expansion for $\phi(t, \nu)$ is the integration by parts formula (4.2), just as (3.1) is the outer expansions for $F(t, \nu)$. In particular,

$$O_r \phi(t, \nu) = \nu^{-1} g(t, t)/h^{[1]}(t, t). \tag{4.5}$$

In addition, we have

$$\phi(\varepsilon T, \nu) = \varepsilon^\alpha e^{T^r \eta(\varepsilon T, \varepsilon T)} \int_T^\infty e^{-X^r \eta(\varepsilon X, \varepsilon T)} g(\varepsilon X, \varepsilon T) X^{\alpha-1} \, dX, \tag{4.6}$$

where $\eta(x, t) = x^{-r} h(x, t)$ and $\varepsilon = \nu^{-1/r}$. The inner expansion for $\phi(t, \nu)$ is obtained directly from here if we just expand in powers of ε, as if X and T were fixed, and then integrate term by term.

A simple example will show how this goes. Let $h(x, t) = x^3(1+t+3x^2)$, $g(x, t) = 1$ and $\alpha = 1$. Substitution into (4.5) yields $O_3\phi(t, \nu) = \nu^{-1}[3t^2(1+t+5t^2)]^{-1}$. It follows that $[O_3 - I_3O_3]\phi(t, \nu) = \nu^{-1}U(t)$, where

$$U(t) = (3t+\tfrac{20}{3}t^2)/(1+t+5t^2). \tag{4.7}$$

In addition, with $\eta(x, t) = 1+t+3x^2$ in (4.6) we can readily calculate $I_3\phi(t, \nu)$. Indeed, for fixed $T \neq \infty$,

$$\int_T^\infty e^{-X^3\eta(\varepsilon X, \varepsilon T)}\, dX = \int_T^\infty [1 - \varepsilon TX^3 - \tfrac{1}{2}\varepsilon^2(6X^5 - T^2X^6)]\, e^{-X^3}\, dX + O(\varepsilon^3), \tag{4.8}$$

and thus, for this example,

$$\phi(t, \nu) = \nu^{-1/3}V_0(\nu^{1/3}t) + \nu^{-2/3}V_1(\nu^{1/3}t) + \nu^{-1}[U(t) + V_2(\nu^{1/3}t)] + O(\nu^{-4/3}) \tag{4.9}$$

uniformly for $0 \leqq t < \infty$, where

$$V_0(T) = e^{T^3}\int_T^\infty e^{-X^3}\, dX, \tag{4.10}$$

$$V_1(T) = (T^4 - \tfrac{1}{3}T)V_0(T) - \tfrac{1}{3}T^2, \tag{4.11}$$

$$V_2(T) = (\tfrac{2}{9}T^2 + \tfrac{8}{3}T^5 + \tfrac{1}{2}T^8)V_0(T) - (1 + \tfrac{7}{9}T^3 + \tfrac{1}{6}T^6). \tag{4.12}$$

Our methods are applicable also to integrals of the form

$$\psi(t, \nu) = e^{\nu h(t,t)}\int_0^\infty e^{-\nu h(x,t)}g(x, t)x^{\alpha-1}\, dx. \tag{4.13}$$

To illustrate, we shall take $h(x, t) = \sinh x \operatorname{sech} t - x$ and $g(x, t) = 1$. Thus $\psi(t, \nu) = A(t, \nu)$, where $A(t, \nu)$ is given by (1.2). The idea here is that the exponent $h(x, t)$ has its minimum at $x = t$, as in (1.1) and (4.1), but now $h^{[1]}(t, t) = 0$. Instead of $h^{[1]}(t, t) > 0$ for $t > 0$, now we have $h^{[2]}(t, t) = \frac{1}{2}\tanh t > 0$ for $t > 0$, but $h^{[2]}(0, 0) = 0$.

Let $c(t) = 6t^{-1}h^{[2]}(t, t)$. By means of Corollary 1, since $c(0) > 6h^{[3]}(0, 0) = 1$, we can show (by dividing the integral into two parts, $0 \leqq x \leqq t$ and $x \geqq t$) there exists $b > 0$ such that

$$A(t, \nu) = \nu^{-\alpha/3}\sum_{k=0}^{N-1}\nu^{-k/3}Q_k(t, \nu^{1/3}t) + O(\nu^{-(N+1)/3}) \tag{4.14}$$

uniformly for $0 \leqq t \leqq b$, where

$$Q_k(t, T) = \int_{-T}^\infty X^k q_k(X, t)\, e^{-(X^3 + c(t)TX^2)/6}(X+T)^{\alpha-1}\, dX, \tag{4.15}$$

and $q_k(X, t)$ is a polynomial of degree k in X^3. Also,

$$Q_k(t, T) \sim T^{\alpha-3/2}\sum_{n=0}^\infty c_{kn}(t)T^{-3n/2}, \qquad T \to \infty, \tag{4.16}$$

in analogy with (3.16), and thus we can apply Theorem 2 to $f(s, S) = Q_k(s^2, S^2)$. Therefore $A(t, \nu)$ has matching inner and outer expansions, and a uniformly valid composite expansion.

The leading term of the outer expansion for $A(t, \nu)$ can be determined directly by the standard Laplace method. Indeed,

$$A(t, \nu) = \nu^{-1/2}y_0(t) + O(\nu^{-3/2}), \qquad t \neq 0, \tag{4.17}$$

where, $y_0(t) = t^{\alpha-1}(2\pi/\tanh t)^{1/2}$. To get the inner expansion observe that

$$A(\varepsilon T, \nu) = \varepsilon^{\alpha} \int_{-T}^{\infty} e^{-\nu\psi(\varepsilon X, \varepsilon T)}(X+T)^{\alpha-1}\, dX \tag{4.18}$$

where $\psi(x, t) = \sinh(x+t)\operatorname{sech} t - x - \tanh t$ and $\varepsilon = \nu^{-1/3}$. Since

$$\psi(x, t) = \tfrac{1}{6}(x^3+3tx^2) + \tfrac{1}{24}(tx^4 - 4t^3x^2) + O((x^2+t^2)^{7/2}), \tag{4.19}$$

it follows that

$$A(\nu^{-1/3}T, \nu) = \nu^{-\alpha/3}[Y_0(T) + \nu^{-2/3}Y_1(Y) + O(\nu^{-4/3})] \tag{4.20}$$

for $T \neq \infty$, where

$$Y_0(T) = \int_{-T}^{\infty} e^{-(X^3+3TX^2)/6}(X+T)^{\alpha-1}\, dX, \tag{4.21}$$

$$Y_1(T) = \frac{1}{24}\int_{-T}^{\infty} e^{-(X^3+3TX^2)/6}(TX^4 - 4T^3X^2)(X+T)^{\alpha-1}\, dX. \tag{4.22}$$

In addition,

$$\nu^{-1/2}y_0(\varepsilon T) = (2\pi)^{1/2}\nu^{-\alpha/3}T^{\alpha-3/2}[1 + \tfrac{1}{6}\nu^{-2/3}T^2 + O(\nu^{-4/3})]; \tag{4.23}$$

thus, combining (4.17), (4.20), and (4.23), we obtain the uniformly valid composite formula (1.3).

REFERENCES

[1] J. S. ANGELL AND W. E. OLMSTEAD, *Singularly perturbed Volterra integral equations*, SIAM J. Appl. Math., 47 (1987), pp. 1150-1162.

[2] N. BLEISTEIN, *Uniform asymptotic expansions of integrals with many nearby stationary points and algebraic singularities*, J. Math. Mech., 17 (1967), pp. 533-559.

[3] C. CHESTER, B. FRIEDMAN, AND F. URSELL, *An extension of the method of steepest descents*, Proc. Camb. Phil. Soc., 54 (1957), pp. 599-611.

[4] A. ERDELYI, *Asymptotic evaluation of integrals involving a fractional derivative*, SIAM J. Math. Anal., 5 (1974), pp. 159-171.

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

[6] L. A. SKINNER, *Uniformly valid expansions for Laplace integrals*, SIAM J. Math. Anal., 11 (1980), pp. 1058-1067.

[7] ———, *Asymptotic evaluation of integrals involving multiple scales*, J. Math. Anal. Appl., 89 (1982), pp. 203-211.

[8] K. SONI, *A note on uniform asymptotic expansions of incomplete Laplace integrals*, SIAM J. Math. Anal., 14 (1983), pp. 1015-1018.

[9] N. M. TEMME, *Remarks on a paper of A. Erdelyi*, SIAM J. Math. Anal., 7 (1976), pp. 767-770.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

012

A DEFINITE INTEGRAL OF A PRODUCT OF TWO POLYLOGARITHMS*

V. S. ADAMCHIK† AND K. S. KÖLBIG‡

Abstract. Using the product theorem for the Mellin transform, a definite integral depending on several parameters and containing a product of two polylogarithm functions (or two logarithms in the degenerate case) is replaced by a Mellin-Barnes integral, which in turn is evaluated by residue techniques. The results are given in terms of infinite series of hypergeometric type. Some special cases for which the infinite series can be expressed in closed form are also considered.

Key words. polylogarithms, logarithmic integrals, Mellin-Barnes integral, hypergeometric series

AMS(MOS) subject classifications. primary 33A70; secondary 33A35, 44A15

1. Introduction.

It is the purpose of this paper to evaluate the integral

$$I_{n,m}(\alpha, \sigma, \omega, r) = \int_0^\infty x^{\alpha-1} Li_n(-\sigma x) Li_m(-\omega x^r)\, dx \tag{1.1}$$

for positive integers n, m, complex α, σ, ω, and real $r \neq 0$ such that the integral exists. $Li_k(z)$ is the polylogarithm function defined for $|z| \leqq 1$ and $k \geqq 2$ by the power series [9, p. 189]

$$Li_k(z) = \sum_{j=1}^{\infty} \frac{z^j}{j^k} \qquad (|z| \leqq 1) \tag{1.2}$$

and for $|z| > 1$ by [9, p. 192], [8]

$$\begin{aligned} Li_k(z) = (-1)^{k+1} Li_k\left(\frac{1}{z}\right) - \frac{1}{k!} \ln^k(-z) \\ - \sum_{j=0}^{k-2} \frac{1}{j!} (1+(-1)^{k-j})(1-2^{1-k+j}) \zeta(k-j) \ln^j(-z), \end{aligned} \tag{1.3}$$

where $\zeta(q)$ is the Riemann zeta function and where the logarithm is defined on its principal sheet. For $k = 1$ we obtain from (1.2)

$$Li_1(z) = -\ln(1-z) \tag{1.4}$$

as a degenerate case. Interest in polylogarithms, which have been investigated in the past by many mathematicians, has revived in recent years because of their applications in quantum electrodynamics, group theory, and geometry. For example, Berndt and Joshi [2] have analysed Ramanujan's work on these functions, Maier and Kiesewetter [10] have investigated systematically some of their functional relations, and Böhm and Hertel [3] have made use of them in the theory of n-dimensional polyhedrals. Gastmans and Troost [5] and Devoto and Duke [4] present tables of integrals leading to or containing polylogarithms which are useful in quantum electrodynamics.

An integral representation of $Li_k(z)$ given recently by Marichev [12, p. 105], namely

$$Li_k(-z) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} \frac{\Gamma(s)\Gamma(-s)}{(-s)^{k-1}} z^{-s}\, ds \qquad (-1 < a < 0, |\arg z| < \pi, k = 0, 1, 2, \cdots) \tag{1.5}$$

* Received by the editors October 8, 1986; accepted for publication (in revised form) June 18, 1987.
† Belorussian State University, Minsk, U.S.S.R.
‡ European Organization for Nuclear Research (CERN), CH-1211 Geneva 23, Switzerland.

allows us to evaluate the integral (1.1) by using the product theorem of the Mellin transform, and by applying the residue theorem.

From (1.3), we find that

$$Li_k(-z) = -\frac{1}{k!}\ln^k z + O(\ln^{k-2} z) \qquad (z \to \infty) \tag{1.6}$$

which can be used, together with (1.2), to show that the integral (1.1) converges under the conditions

$$\begin{aligned} &-1-r<\operatorname{Re}\alpha<0 \quad \text{if } r>0, \\ &-1<\operatorname{Re}\alpha<-r \quad \text{if } r<0, \end{aligned} \tag{1.7}$$

and $|\arg\sigma|<\pi$, $|\arg\omega|<\pi$.

We may add here that, by making the substitution $x=\xi^{1/r}$, we obtain from (1.1) the "symmetry" relation

$$I_{n,m}(\alpha,\sigma,\omega,r) = \frac{1}{|r|} I_{m,n}\left(\frac{\alpha}{r}, \omega, \sigma, \frac{1}{r}\right). \tag{1.8}$$

2. An alternative integral for $I_{n,m}(\alpha,\sigma,\omega,r)$. In this section, we derive a Mellin-Barnes integral for $I_{n,m}$. By introducing (1.5) into (1.1), we find with the (allowed) substitution $rs=-s'$,

$$\begin{aligned} I_{n,m} = \frac{1}{\sigma^\alpha |r|}\int_0^\infty \frac{du}{u}\left(u^\alpha \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \frac{\Gamma(s)\Gamma(-s)}{(-s)^{n-1}} u^{-s}\,ds\right) \\ \cdot\left(\frac{1}{2\pi i}\int_{a'-i\infty}^{a'+i\infty} \frac{\Gamma(s/r)\Gamma(-s/r)}{(s/r)^{m-1}}\left(\frac{\zeta}{u}\right)^{-s} ds\right) \end{aligned}$$

where $\zeta=\sigma\omega^{-1/r}$. Applying well-known theorems for the Mellin transform (see, e.g., Sneddon [17, pp. 262-297]), in particular the product theorem

$$\int_0^\infty f\left(\frac{\zeta}{u}\right) g(u)\frac{du}{u} = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} f^*(s)g^*(s)\zeta^{-s}\,ds,$$

we see that

$$\begin{aligned} g^*(s) &= (-1)^{n+1}\Gamma(s+\alpha)\Gamma(-s-\alpha)(s+\alpha)^{1-n}, \\ f^*(s) &= \Gamma\left(\frac{s}{r}\right)\Gamma\left(-\frac{s}{r}\right)\left(\frac{s}{r}\right)^{1-m}, \end{aligned}$$

and therefore that

$$I_{n,m}(\alpha,\sigma,\omega,r) = \frac{(-1)^{n+1}}{\sigma^\alpha|r|}\frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \frac{\Gamma(s+\alpha)\Gamma(-s-\alpha)\Gamma(s/r)\Gamma(-s/r)}{(s+\alpha)^{n-1}(s/r)^{m-1}}\zeta^{-s}\,ds, \tag{2.1}$$

where

$$\begin{aligned} &\max(0,-1-\operatorname{Re}\alpha)<\gamma=\operatorname{Re}s<\min(r,-\operatorname{Re}\alpha) \qquad (r>0), \\ &\max(r,-1-\operatorname{Re}\alpha)<\gamma=\operatorname{Re}s<\min(0,-\operatorname{Re}\alpha) \qquad (r<0). \end{aligned} \tag{2.2}$$

Using [6, No. 8.3343]

$$\Gamma(z)\Gamma(-z) = -\frac{\pi}{z}\csc\pi z,$$

the integral (2.1) can be written as

$$I_{n,m}(\alpha, \sigma, \omega, r) = \frac{(-1)^{n+1}\pi^2}{\sigma^{\alpha}|r|}\frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \csc[\pi(s+\alpha)]\csc\left(\pi\frac{s}{r}\right)(s+\alpha)^{-n}\left(\frac{s}{r}\right)^{-m}\zeta^{-s}\,ds. \tag{2.3}$$

Formula (2.3) is more suitable than (2.1) for evaluation by the residue theorem.

We may note here that $I_{n,m}$ is a special case of Fox's H-function (see Mathai and Saxena [14, p. 2]), namely

$$I_{n,m}(\alpha, \sigma, \omega, r) = \frac{1}{\sigma^{\alpha}|r|}H^{m+1,n+1}_{m+n,m+n} \times\left[\frac{\sigma}{\omega^{1/r}}\middle|\begin{matrix}(1+\alpha,1),\cdots,(1+\alpha,1),(1,1/r),\cdots,(1,1/r)\\ (0,1/r),\cdots,(0,1/r),(\alpha,1),\cdots,(\alpha,1)\end{matrix}\right] \tag{2.4}$$

with n repetitions of the brackets $(1+\alpha, 1)$, $(\alpha, 1)$ and m repetitions of $(1, 1/r)$, $(0, 1/r)$.

For rational $r = \pm p/q$, $(p, q \in \mathbb{N})$ it is possible to express the integral (2.1) as a special case of Meijer's G-function in the logarithmic case [13, p. 176]. Substituting $s' = s/p$ in (2.1), and applying the product theorem [6, No. 8.335] of the gamma function we obtain

$$I_{n,m}\left(\alpha, \sigma, \omega, \pm\frac{p}{q}\right) = (-1)^{n+1}(\pm 1)^{m+1}\frac{(2\pi)^{2-p-q}}{\sigma^{\alpha}p^{n}q^{m-1}}G^{m',n'}_{p',p'}\left(\frac{\sigma^{p}}{\omega^{\pm q}}\middle|\begin{matrix}a_j\\ b_j\end{matrix}\right), \tag{2.5}$$

where

$$\begin{aligned} m' &= n+m+p+q-2,\\ n' &= p+q,\\ p' &= n+m+p+q-2, \end{aligned} \tag{2.6}$$

and

$$a_j = \begin{cases} 1+(1+\alpha-j)/p, & j=1,\cdots,p,\\ 1+(1-j)/q, & j=p+1,\cdots,p+q,\\ 1+\alpha/p, & j=p+q+1,\cdots,p+q+n-1,\\ 1, & j=p+q+n,\cdots,p+q+n+m-2, \end{cases}$$

$$b_j = \begin{cases} \alpha/p, & j=1,\cdots,n-1,\\ 0, & j=n,\cdots,n+m-2,\\ (\alpha+j-1)/p, & j=n+m-1,\cdots,n+m+p-2,\\ (j-1)/q, & j=n+m+p-1,\cdots,n+m+p+q-2. \end{cases} \tag{2.7}$$

According to the theory of the G-function [15], we know that $G^{m',n'}_{p',p'}(z)$ is analytic and continuous in the whole sector $|\arg z| < \pi(m'+n'-p')$, provided that $m'+n'-p' > 1$. From (2.6), we have that $m'+n'-p' = p+q > 1$, which further implies [15] that the point $(-1)^{m'+n'-p'}$ is not a singular point. It follows that, for rational r, the function (2.1) is analytic and continuous in the whole sector $|\arg \zeta| < \pi(1+q/p)$. Using the analyticity of the integral (2.1) with respect to r, it follows that this property holds for any real r. We shall use this continuity property of (2.1) when considering the examples in § 4.

3. The evaluation of the integral. Using techniques developed by Marichev [11], [12] and, for the logarithmic case of integrals (2.1), by Adamchik and Marichev [1] and Mathai and Saxena [13, p. 157], [14, p. 70], it is possible to evaluate the integral (1.1) in terms of infinite (hypergeometric type) series. We thus obtain the following theorem.

THEOREM. *Let $r \neq 0$ be real, α, σ, ω be complex with $|\arg \sigma| < \pi$, $|\arg \omega| < \pi$, and let $\zeta = \sigma\omega^{-1/r}$. Further, with $\mathbb{N} = \{1, 2, 3, \cdots\}$, let $\mathbb{N}_0 = \{0, \mathbb{N}\}$,*

$$h(x) = \begin{cases} 1 & \text{if } x \in \mathbb{N}, \\ 0 & \text{otherwise,} \end{cases} \qquad \bar{h}(x) = 1 - h(x);$$

$$H(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x < 0, \end{cases}$$

and let k, l, $K \in \mathbb{N}$.

Then, for $|\zeta| \neq 1$ and

$$-1 - r < \operatorname{Re} \alpha < 0 \quad \text{if } r > 0,$$

$$-1 < \operatorname{Re} \alpha < -r \quad \text{if } r < 0,$$

the integral (1.1) *can, for $\alpha \neq 0$, be expressed in the form*

(3.1)

$$\begin{aligned}
&\int_0^\infty x^{\alpha-1} Li_n(-\sigma x) Li_m(-\omega x^r)\, dx \\
&\quad = (-1)^n (\pm \operatorname{sgn} r)^{m+1} \pi \sigma^{-\alpha} \sum_{|r|k \pm \alpha \notin \mathbb{N}_0} (-1)^k \csc[\pi(\alpha \pm |r|k)] (\alpha \pm |r|k)^{-n} k^{-m} \zeta^{\mp |r| k} \\
&\qquad \mp (\pm 1)^n (-1)^{m+n} \frac{\pi}{|r|} \omega^{-\alpha/r} \sum_{(l \mp \alpha)/|r| \notin \mathbb{N}_0} (-1)^l \csc\left(\pi \frac{\alpha \mp l}{r}\right) \left(\frac{\alpha \mp l}{r}\right)^{-m} l^{-n} \zeta^{\mp l} \\
&\qquad + (-1)^{n+1} (\pm \operatorname{sgn} r)^{m+1} \sigma^{-\alpha} \sum_{|r|K \pm \alpha \in \mathbb{N}} (-1)^{K + |r|K \pm \alpha} (\alpha \pm |r|K)^{-n} K^{-m} \zeta^{\mp |r| K} \\
&\qquad \cdot \left(\pm \frac{m}{|r|K} + \frac{n}{\alpha \pm |r|K} + \ln \zeta \right) + H(\mp r)(-1)^{n+1} \sigma^{-\alpha} \pi^{m+1} r^m \alpha^{-n} \\
&\qquad \cdot \Bigg\{ h(\pm\alpha)(-1)^\alpha \sum_{m_1=0}^{m+1} \frac{1}{m_1!} \left(-\frac{1}{\pi} \ln \zeta \right)^{m_1} \\
&\qquad\quad \cdot \sum_{m_2=0}^{m+1-m_1} \binom{m_2 + n - 1}{n-1} (-\pi\alpha)^{-m_2} \sum_{m_3=0}^{m+1-m_1-m_2} r^{-m_3} B^*_{m_3} B^*_{m+1-m_1-m_2-m_3} \\
&\qquad\quad + \bar{h}(\pm\alpha) \sum_{m_1=0}^{m} \frac{1}{m_1!} \left(-\frac{1}{\pi} \ln \zeta \right)^{m_1} \\
&\qquad\quad \cdot \sum_{m_2=0}^{m-m_1} \binom{m_2 + n - 1}{n-1} (-\pi\alpha)^{-m_2} \sum_{m_3=0}^{m-m_1-m_2} r^{-m_3} B^*_{m_3} C_{m-m_1-m_2-m_3}(\pi\alpha) \Bigg\} \\
&\qquad + H(\pm 1)(-1)^m (\operatorname{sgn} r) \omega^{-\alpha/r} \pi^{n+1} r^m \alpha^{-m} \\
&\qquad \cdot \Bigg\{ h\left(-\frac{\alpha}{|r|}\right)(-1)^{n+\alpha/r} \sum_{n_1=0}^{n+1} \frac{1}{n_1!} \left(-\frac{1}{\pi} \ln \zeta \right)^{n_1} \\
&\qquad\quad \cdot \sum_{n_2=0}^{n+1-n_1} \binom{n_2 + m - 1}{m-1} (\pi\alpha)^{-n_2} \sum_{n_3=0}^{n+1-n_1-n_2} r^{-n_3} B^*_{n_3} B^*_{n+1-n_1-n_2-n_3}
\end{aligned}$$

$$-\bar{h}\left(-\frac{\alpha}{|r|}\right)r^{-n-1}\sum_{n_1=0}^{n}\frac{1}{n_1!}\left(\frac{r}{\pi}\ln\zeta\right)^{n_1}$$
$$\cdot\sum_{n_2=0}^{n-n_1}(-1)^{n_2}\binom{n_2+m-1}{m-1}\left(\pi\frac{\alpha}{r}\right)^{-n_2}\sum_{n_3=0}^{n-n_1-n_2}(-r)^{n_3}B_{n_3}^{*}C_{n-n_1-n_2-n_3}\left(\pi\frac{\alpha}{r}\right)\Bigg\},$$

and for $\alpha=0$ *(which implies* $r<0$*), in the form*

$$\int_0^\infty Li_n(-\sigma x)Li_m(-\omega x^r)\frac{dx}{x}$$
$$=-(\mp1)^{m+n}\Bigg\{|r|^{-n}\pi\sum_{|r|k\notin\mathbb{N}}(-1)^k\csc(\pi|r|k)k^{-(m+n)}\zeta^{\mp|r|k}$$
$$+|r|^{m-1}\pi\sum_{l/|r|\notin\mathbb{N}}(-1)^l\csc\left(\pi\frac{l}{|r|}\right)l^{-(m+n)}\zeta^{\mp l}$$
(3.2)
$$-|r|^{-n}\sum_{|r|K\in\mathbb{N}}(-1)^{K(1+r)}K^{-(m+n)}\zeta^{\mp|r|K}\left(\frac{m+n}{|r|K}\pm\ln\zeta\right)\Bigg\}$$
$$+H(\pm1)(-1)^{n+1}\pi^{m+n+1}r^m$$
$$\cdot\sum_{m_1=0}^{m+n+1}\frac{1}{m_1!}\left(-\frac{1}{\pi}\ln\zeta\right)^{m_1}\sum_{m_2=0}^{m+n+1-m_1}r^{-m_2}B_{m_2}^{*}B_{m+n+1-m_1-m_2}^{*},$$

where

$$B_j^{*}=\frac{|2^j-2|}{j!}|B_j|, \tag{3.3}$$

B_j *are the Bernoulli numbers, where*

$$C_j(x)=\frac{1}{j!}\frac{d^j}{ds^j}\csc(s+x)|_{s=0}, \tag{3.4}$$

and where the logarithm is defined on its principal sheet.

In all the expressions above, the upper sign corresponds to $|\zeta|>1$, the lower sign to $|\zeta|<1$.

Proof. Using the theory of the H-function [14], we find that the poles of the integrand in (2.3), i.e., of

$$\varphi_{n,m}(s;\alpha,r)=\csc[\pi(s+\alpha)]\csc\left(\pi\frac{s}{r}\right)(s+\alpha)^{-n}\left(\frac{s}{r}\right)^{-m}, \tag{3.5}$$

which need to be taken into account when evaluating the integral (1.1), are those which, for $|\zeta|>1$ lie on the right, and for $|\zeta|<1$ on the left, of the line $\operatorname{Re} s=\gamma$. In the first case the values of the corresponding residues must be multiplied by a factor -1. The relevant poles are:

$|\zeta|>1$.

(i) At $s=0$ if $r<0$. Pole of order $m+2$ if $\alpha\in\mathbb{N}$, of order $m+n+2$ if $\alpha=0$, of order $m+1$ if $\alpha\notin\mathbb{N}_0$.

(ii) At $s=-\alpha$. If $r<0$: pole of order $n+2$ if $\alpha/r\in\mathbb{N}$, of order $n+1$ if $\alpha/r\notin\mathbb{N}_0$. If $r>0$: pole of order $n+2$ if $-\alpha/r\in\mathbb{N}$, of order $n+1$ if $-\alpha/r\notin\mathbb{N}$.

(iii) At $s=|r|k$ and at $s=-\alpha+l$, for $k,l\in\mathbb{N}$. Both poles of order 1 if $|r|k\neq-\alpha+l$, single pole of order 2 if $|r|k=-\alpha+l$.

$|\zeta|<1$.

(i) At $s=0$ if $r>0$. Pole of order $m+2$ if $-\alpha\in\mathbb{N}$, of order $m+1$ if $-\alpha\notin\mathbb{N}$.

(ii) At $s=-|r|k$ and at $s=-\alpha-l$ for $k,l\in\mathbb{N}$. Both poles of order 1 if $|r|k\neq\alpha+l$, single pole of order 2 if $|r|k=\alpha+l$.

The residues corresponding to these poles can be evaluated by standard methods, making use of the Leibniz formula for the mth derivative of a product of k functions of one variable, together with elementary properties of the cosecant and cotangent functions, and the power series expansion [6, No. 1.41111]

$$\pi x \csc \pi x = \sum_{j=0}^{\infty} \frac{|2^j - 2|}{j!} |B_j| (\pi x)^j \qquad (|x| < 1). \tag{3.6}$$

By collecting the computed residues appropriately we obtain formulae (3.1) and (3.2) of the theorem.

We here list, for convenience, the first six functions $C_j(x)$ defined by (3.4):

$$C_j(x) = \frac{1}{j!} \frac{d^j}{ds^j} \csc(s+x)\Big|_{s=0} = \frac{1}{j!} \csc^{j+1} x \tilde{C}_j(x) \tag{3.7}$$

where

$$\tilde{C}_0(x) = 1,$$
$$\tilde{C}_1(x) = -\cos x,$$
$$\tilde{C}_2(x) = 2 - \sin^2 x,$$
$$\tilde{C}_3(x) = \cos x(\sin^2 x - 6),$$
$$\tilde{C}_4(x) = \sin^4 x - 20 \sin^2 x + 24,$$
$$\tilde{C}_5(x) = \cos x(-\sin^4 x + 60 \sin^2 x - 120).$$

We may add here that it is not difficult to evaluate the finite sums in (3.1) and (3.2) by a symbolic algebra system such as REDUCE [7].

4. Special cases. In this section, we consider some special cases of (3.1) and (3.2). For notational purposes, we make use of the Lerch function [6, No. 9.550]

$$\Phi(z, u, v) = \sum_{k=0}^{\infty} \frac{z^k}{(k+v)^u} \qquad (|z| < 1, -v \notin \mathbb{N}_0), \tag{4.1}$$

which is a generalization of the polylogarithm function (1.2). In particular, for $u = 2$, $v = \frac{1}{2}$, we have

$$\Phi\left(-z, 2, \frac{1}{2}\right) = \frac{4}{\sqrt{z}} \sum_{k=0}^{\infty} \frac{(-1)^k (\sqrt{z})^{2k+1}}{(2k+1)^2} = \frac{4}{\sqrt{z}} Ti_2(\sqrt{z}) \tag{4.2}$$

where

$$Ti_2(z) = \int_0^z t^{-1} \arctan t \, dt$$

is the arctangent integral.

We shall also make use of the following lemma.

LEMMA. *Let E_j be the Euler numbers* [6, No. 9.72], *let*

$$E_j^* = \begin{cases} 0, & j \in \{-2, -1\}, \\ |E_j|/j!, & j \in \mathbb{N}_0, \end{cases}$$

and let B_j^* be defined by (3.3). *Then*

$$S_k \equiv \sum_{j=0}^{k} (\pm 2)^{k-j} B_j^* B_{k-j}^* = E_{k-2}^* - (k-1) B_k^* \qquad (k \in \mathbb{N}_0). \tag{4.3}$$

Proof. We start from the identity

$$(x \csc x)(2x \csc 2x) = x^2\left(\sec x - \frac{d}{dx}\csc x\right)$$

and replace the trigonometric functions by their power series (3.6) and [6, No. 1.4119]

$$\sec x = \sum_{j=0}^{\infty} \frac{|E_j|}{j!} x^j \qquad \left(|x| < \frac{1}{2}\pi\right),$$

respectively. Multiplying together the two power series on the left-hand side of the identity and equating coefficients yields the lemma.

4.1. $r=-2$, $\alpha=0$. In this case the summation conditions for the infinite series become

$$2k \notin \mathbb{N}, \quad \tfrac{1}{2}l \notin \mathbb{N}, \quad 2K \in \mathbb{N},$$

which exclude all $k \in \mathbb{N}$ but permit all odd values of $l \in \mathbb{N}$, and all $K \in \mathbb{N}$. Thus, from (3.2) with $\zeta = \sigma\sqrt{\omega}$, using (1.2) and (4.3),

(4.4)

$$\begin{aligned}
&\int_0^\infty Li_n(-\sigma x) Li_m(-\omega x^{-2}) \frac{dx}{x} \\
&\quad = (\mp)^{m+n}\left\{2^{-(n+1)}\pi\zeta^{\mp 1}\Phi\left(-\zeta^{\mp 2}, m+n, \frac{1}{2}\right)\right. \\
&\qquad\qquad \left. + 2^{-(n+1)}(m+n) Li_{m+n+1}(-\zeta^{\mp 2}) \pm 2^{-n} \ln\zeta\, Li_{m+n}(-\zeta^{\mp 2})\right\} \\
&\quad + H(\pm 1)(-\pi)^{m+n+1} 2^{-n-1} \sum_{j=0}^{m+n+1} \frac{1}{j!}\left(-\frac{2}{\pi}\ln\zeta\right)^j S_{m+n+1-j} \\
&\qquad\qquad (|\zeta| \gtrless 1,\ m, n \in \mathbb{N},\ |\arg\sigma| < \pi,\ |\arg\omega| < \pi).
\end{aligned}$$

In particular, for $n = m = 1$, using (1.4) and (4.2):

$$\begin{aligned}
\int_0^\infty \ln(1+\sigma x)\ln(1+\omega x^{-2})\frac{dx}{x} &= \pi Ti_2(\zeta^{\mp 1}) + \frac{1}{2} Li_3(-\zeta^{\mp 2}) \pm \frac{1}{2}\ln\zeta Li_2(-\zeta^{\mp 2}) \\
&\quad + \begin{cases} \frac{1}{3}\ln^3\zeta + \frac{5}{12}\pi^2\ln\zeta & (|\zeta| > 1), \\ 0 & (|\zeta| < 1). \end{cases}
\end{aligned} \tag{4.5}$$

For $n = 1$, $m = 2$:

$$\begin{aligned}
\int_0^\infty \ln(1+\sigma x) Li_2(-\omega x^{-2})\frac{dx}{x} &= \pm\frac{1}{4}\pi\zeta^{\mp 1}\Phi\left(-\zeta^{\mp 2}, 3, \frac{1}{2}\right) \pm \frac{3}{4} Li_4(-\zeta^{\mp 2}) \\
&\quad + \frac{1}{2}\ln\zeta\, Li_3(-\zeta^{\mp 2}) \\
&\quad - \begin{cases} \frac{1}{6}\ln^4\zeta + \frac{5}{12}\pi^2\ln^2\zeta + \frac{53}{480}\pi^4 & (|\zeta| > 1), \\ 0 & (|\zeta| < 1). \end{cases}
\end{aligned} \tag{4.6}$$

For $n=2$, $m=1$:

$$\int_0^\infty Li_2(-\sigma x)\ln(1+\omega x^{-2})\frac{dx}{x}=\frac{1}{2}\int_0^\infty \ln(1+\sigma x)Li_2(-\omega x^{-2})\frac{dx}{x}. \tag{4.7}$$

For $n=m=2$:

$$\int_0^\infty Li_2(-\sigma x)Li_2(-\omega x^{-2})\frac{dx}{x}=\frac{1}{8}\pi\zeta^{\mp1}\Phi\left(-\zeta^{\mp2},4,\frac{1}{2}\right)+\frac{1}{2}Li_5(-\zeta^{\mp2})\pm\frac{1}{4}\ln\zeta Li_4(-\zeta^{\mp2})$$

$$-\begin{cases}\dfrac{1}{30}\ln^5\zeta+\dfrac{5}{36}\pi^2\ln^3\zeta+\dfrac{53}{480}\pi^4\ln\zeta & (|\zeta|>1),\\ 0 & (|\zeta|<1).\end{cases} \tag{4.8}$$

Using the fact that the integral is continuous for $|\arg\sigma\sqrt{\omega}|<\pi/2$, we can set $\sigma=\omega=1$ and obtain from (4.4):

$$\int_0^\infty Li_n(-x)Li_m(-x^{-2})\frac{dx}{x}=2^{-(n+1)}\left\{\pi\Phi\left(-1,n+m,\frac{1}{2}\right)+(n+m)Li_{n+m+1}(-1)\right\}$$

$$=2^{-(n+1)}\left\{\pi 2^{n+m}\sum_{k=0}^{\infty}\frac{(-1)^k}{(2k+1)^{n+m}}-(n+m)\sum_{k=1}^{\infty}\frac{(-1)^k}{k^{n+m+1}}\right\}. \tag{4.9}$$

In the case of $n+m$ odd, these series are well known (e.g., [16, No. 5.1.3.3, 5.1.4.2]). Thus we obtain

$$\int_0^\infty Li_n(-x)Li_m(-x^{-2})\frac{dx}{x}=2^{-(n+2)}\frac{\pi^{n+m+1}}{(n+m-1)!}\left\{|E_{n+m-1}|-\frac{2^{n+m+1}-2}{n+m+1}|B_{n+m+1}|\right\} \tag{4.10}$$

$(n, m\in\mathbb{N}, n+m \text{ odd}).$

This formula is remarkable insofar as it contains both the Bernoulli and Euler numbers. For $n=m=1$, formula (4.9) reduces to

$$\int_0^\infty \ln(1+x)\ln(1+x^{-2})\frac{dx}{x}=\pi G-\frac{3}{8}\zeta(3), \tag{4.11}$$

where G is Catalan's constant.

4.2. $r=2$, $\alpha=-2$. In this case the summation conditions for the infinite series become

$$2k\mp2\notin\mathbb{N}_0,\quad \tfrac{1}{2}l\pm1\notin\mathbb{N}_0,\quad 2K\mp2\in\mathbb{N},$$

which exclude all $k\in\mathbb{N}$, but permit all odd values of $l\in\mathbb{N}$, all $K\in\mathbb{N}\setminus\{1\}$ if $|\zeta|>1$, and all $K\in\mathbb{N}$ if $|\zeta|<1$. Thus, from (3.1) with $\zeta=\sigma/\sqrt{\omega}$, using (4.3),

$$\int_0^\infty x^{-3}Li_n(-\sigma x)Li_m(-\omega x^2)dx$$

$$=(-1)^n\left\{2^{m-1}\pi\frac{\omega^{3/2}}{\sigma}\sum_{l=0}^{\infty}\frac{(-\zeta^{-2})^l}{(2l+1)^n(2l+3)^m}\right.$$

$$+2^{n-1}\omega\left[m\sum_{k=1}^{\infty}\frac{(-\zeta^{-2})^k}{k^n(k+1)^{m+1}}+n\sum_{k=1}^{\infty}\frac{(-\zeta^{-2})^k}{k^{n+1}(k+1)^m}+2\ln\zeta\sum_{k=1}^{\infty}\frac{(-\zeta^{-2})^k}{k^n(k+1)^m}\right] \tag{4.12a}$$

$$\left.-\left(\frac{\pi}{2}\right)^{n+1}\omega\sum_{n_1=0}^{n+1}\frac{1}{n_1!}\left(-\frac{2}{\pi}\ln\zeta\right)^{n_1}\sum_{n_2=0}^{n+1-n_1}\binom{n_2+m-1}{m-1}(-\pi)^{-n_2}S_{n+1-n_1-n_2}\right\}$$

$(|\zeta|>1, m, n\in\mathbb{N}, |\arg\sigma|<\pi, |\arg\omega|<\pi),$

(4.12b)

$$\int_0^\infty x^{-3} Li_n(-\sigma x) Li_m(-\omega x^2)\, dx$$

$$= (-1)^m \left\{ 2^{m-1} \pi \sigma \sqrt{\omega} \sum_{l=0}^{\infty} \frac{(-\zeta^2)^l}{(2l+1)^n (2l-1)^m} - 2^{-n-1} \sigma^2 \right.$$

$$\left. \times \left[m \sum_{k=1}^{\infty} \frac{(-\zeta^2)^k}{k^{m+1}(k+1)^n} + n \sum_{k=1}^{\infty} \frac{(-\zeta^2)^k}{k^m (k+1)^{n+1}} - 2 \ln \zeta \sum_{k=1}^{\infty} \frac{(-\zeta^2)^k}{k^m (k+1)^n} \right] \right\}$$

$$- \pi^{m+1} 2^{-n-1} \sigma^2 \sum_{m_1=0}^{m+1} \frac{1}{m_1!} \left(-\frac{2}{\pi} \ln \zeta \right)^{m_1} \sum_{m_2=0}^{m+1-m_1} \binom{m_2 - n - 1}{n-1} \pi^{-m_2} S_{m+1-m_1-m_2}$$

$$(|\zeta| < 1,\ m, n \in \mathbb{N},\ |\arg \sigma| < \pi,\ |\arg \omega| < \pi).$$

Using the relations [16, Nos. 5.2.5.5,15, 5.2.6.1,2]

$$\sum_{l=0}^{\infty} \frac{(-1)^l x^{2l+3}}{(2l+1)(2l+3)} = \frac{1}{2}(1+x^2) \arctan x - \frac{1}{2} x,$$

$$\sum_{k=1}^{\infty} \frac{x^{k+1}}{k(k+1)^2} = 2x + (1-x) \ln (1-x) - Li_2(x),$$

(4.13)

$$\sum_{k=1}^{\infty} \frac{x^{k+1}}{k^2(k+1)} = (x-1) \ln (1-x) + x(Li_2(x) - 1),$$

$$\sum_{k=1}^{\infty} \frac{x^{k+1}}{k(k+1)} = x + (1-x) \ln (1-x),$$

we find for $n = m = 1$, after some calculation,

(4.14)

$$\int_0^\infty x^{-3} \ln (1+\sigma x) \ln (1+\omega x^2)\, dx$$

$$= \frac{1}{2} \left\{ (\sigma^2 + \omega) \left[\ln \left(\frac{\sigma}{\sqrt{\omega}} \right) \ln \left(1 + \frac{\omega}{\sigma^2} \right) - \frac{1}{2} Li_2 \left(-\frac{\omega}{\sigma^2} \right) - \pi \arctan \left(\frac{\sqrt{\omega}}{\sigma} \right) \right] \right.$$

$$\left. + \omega \left[\pi \frac{\sigma}{\sqrt{\omega}} + \ln^2 \left(\frac{\sigma}{\sqrt{\omega}} \right) + \frac{5}{12} \pi^2 \right] \right\} \qquad (|\sigma/\sqrt{\omega}| > 1),$$

$$= \frac{1}{2} \left\{ (\sigma^2 + \omega) \left[\ln \left(\frac{\sigma}{\sqrt{\omega}} \right) \ln \left(1 + \frac{\sigma^2}{\omega} \right) + \frac{1}{2} Li_2 \left(-\frac{\sigma^2}{\omega} \right) + \pi \arctan \left(\frac{\sigma}{\sqrt{\omega}} \right) \right] \right.$$

$$\left. + \sigma^2 \left[\pi \frac{\sqrt{\omega}}{\sigma} - \ln^2 \left(\frac{\sigma}{\sqrt{\omega}} \right) - \frac{5}{12} \pi^2 \right] \right\} \qquad (|\sigma/\sqrt{\omega}| < 1).$$

For $\sigma \to \sqrt{\omega}$, we obtain in the limit, using $Li_2(-1) = -\pi^2/12$,

(4.15)
$$\int_0^\infty x^{-3} \ln (1+x) \ln (1+x^2)\, dx = \frac{\pi}{2}.$$

Using the fact that the integral is continuous at $\sigma=\omega=1$, we obtain

(4.16)

$$\int_0^\infty x^{-3}Li_n(-x)Li_m(-x^2)\,dx$$
$$=(-1)^n\Bigg\{2^{m-1}\pi\sum_{l=0}^{\infty}\frac{(-1)^l}{(2l+1)^n(2l+3)^m}$$
$$+2^{-n-1}\Bigg[m\sum_{k=1}^{\infty}\frac{(-1)^k}{k^n(k+1)^{m+1}}+n\sum_{k=1}^{\infty}\frac{(-1)^k}{k^{n+1}(k+1)^m}\Bigg]$$
$$-\left(\frac{\pi}{2}\right)^{n+1}\sum_{j=0}^{n+1}\binom{j+m-1}{m-1}(-\pi)^{-j}S_{n+1-j}\Bigg\}.$$

From [16, No. 5.1.24.14] we see that the first sum can in general be expressed only in terms of generalized zeta functions $\zeta(j,\frac{1}{2})$ and $\zeta(j,\frac{1}{4})$. For $n=m$, however, using the relations [16, Nos. 5.1.24.10, 5.1.24.15][1,2]

(4.17)

$$\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k^m(k+1)^n}=(-1)^{m-1}\sum_{k=0}^{n-2}\binom{m+k-1}{k}(1-2^{1-n+k})\zeta(n-k)$$
$$+\sum_{k=0}^{m-2}(-1)^k\binom{n+k-1}{k}(1-2^{1-m+k})\zeta(m-k)$$
$$+(-1)^m\binom{m+n-1}{n-1}+(-1)^{m-1}\binom{m+n-2}{m-1}2\ln 2 \qquad (m,n\in\mathbb{N})$$

and

(4.18)

$$\sum_{l=0}^{\infty}\frac{(-1)^l}{(2l+1)^n(2l+3)^n}$$
$$=(-1)^n\left\{\frac{1}{2}-\frac{\pi}{2^{2n}(n-1)!}\sum_{k=0}^{[(n-1)/2]}\frac{(2n-2k-2)!}{(n-2k-1)!(2k)!}\pi^{2k}|E_{2k}|\right\} \qquad (n\in\mathbb{N})$$

where E_{2k} are the Euler numbers, we find that

(4.19)

$$\int_0^\infty x^{-3}Li_n(-x)Li_n(-x^2)\,dx$$
$$=2^{n-2}\pi-\frac{2^{-n-1}}{(n-1)!}\Bigg\{\sum_{k=0}^{n-1}\frac{(2n-k-2)!}{(n-k-1)!k!}\pi^{k+2}|E_k|$$
$$+\Bigg[\sum_{k=0}^{n-1}[1-(-1)^{n+k}]$$

[1] Note the misprint in [16, No. 5.1.24.10], where the factor $(-1)^m$ reads $(-1)^n$.

[2] Note that [16, No. 5.1.24.15] is erroneous: $|E_{2k}|$ there reads E_{2k} and $(2n-2k-2)!/(n-2k-1)!$ reads $(2n-k-2)!/(n-k-1)!$.

$$\times(k-n)\frac{(n+k-1)!}{k!}(1-2^{k-n})\zeta(n-k+1)$$
$$+\sum_{k=0}^{n}(-1)^{k+n}\frac{(k+n-1)!}{k!}\pi^{n-k+1}S_{n-k+1}\Bigg]\Bigg\}.$$

Because of the fact that only even integers occur as arguments for the zeta function, expression (4.19) represents a polynomial in π with rational coefficients. However, using (4.3) and the relation [6, No. 9.5421]

$$\zeta(2j)=2^{2j-1}\pi^{2j}\frac{|B_{2j}|}{(2j)!} \tag{4.20}$$

it is not difficult to show that the coefficients of all but the first power vanish, and therefore that

$$\int_0^\infty x^{-3}Li_n(-x)Li_n(-x^2)\,dx=2^{n-2}\pi \qquad (n\in\mathbb{N}). \tag{4.21}$$

We note that, for $n=1$, (4.21) agrees with (4.15).

4.3. $r=1$, $\alpha=-1/2$. In this case, the summation conditions for the infinite series become

$$k\mp\tfrac{1}{2}\notin\mathbb{N}_0,\quad l\pm\tfrac{1}{2}\notin\mathbb{N}_0,\quad K\mp\tfrac{1}{2}\in\mathbb{N}$$

which permit all k, $l\in\mathbb{N}$ and exclude all $K\in\mathbb{N}$. Thus from (3.1) with $\zeta=\sigma/\omega$

(4.22)

$$\int_0^\infty x^{-3/2}Li_n(-\sigma x)Li_m(-\omega x)\,dx$$
$$=-(\pm1)^{m+1}2^n\pi\sqrt{\sigma}\sum_{k=1}^{\infty}\frac{\zeta^{\mp k}}{k^m(1\mp2k)^n}\pm(\mp1)^n2^m\pi\sqrt{\omega}\sum_{l=1}^{\infty}\frac{\zeta^{\mp l}}{(1\pm2l)^m l^n}$$
$$-H(\mp1)2^n\pi^{m+1}\sqrt{\sigma}\sum_{m_1=0}^{m}\frac{1}{m_1!}\left(-\frac{1}{\pi}\ln\zeta\right)^{m_1}$$
$$\cdot\sum_{m_2=0}^{m-m_1}\binom{m_2+n-1}{n-1}\left(\frac{2}{\pi}\right)^{m_2}\sum_{m_3=0}^{m-m_1-m_2}B^*_{m_3}C_{m-m_1-m_2-m_3}\left(-\frac{\pi}{2}\right)$$
$$-H(\pm1)2^m\pi^{n+1}\sqrt{\omega}\sum_{n_1=0}^{n}\frac{1}{n_1!}\left(-\frac{1}{\pi}\ln\zeta\right)^{n_1}$$
$$\cdot\sum_{n_2=0}^{n-n_1}\binom{m_2+n-1}{n-1}\left(\frac{2}{\pi}\right)^{n_2}\sum_{n_3=0}^{n-n_1-n_2}(-1)^{n_3}B^*_{n_3}C_{n-n_1-n_2-n_3}\left(-\frac{\pi}{2}\right)$$
$$(|\zeta|\gtrless1,\ m,\ n\in\mathbb{N},\ |\arg\sigma|<\pi,\ |\arg\omega|<\pi).$$

Using the relations

(4.23)

$$\sum_{k=1}^{\infty}\frac{x^{2k}}{k(2k+1)}=2-\frac{2}{x}\operatorname{Arth}x-\ln(1-x^2),$$
$$\sum_{k=1}^{\infty}\frac{x^{2k}}{k(2k-1)}=2x\operatorname{Arth}x+\ln(1-x^2),$$

$$\sum_{k=1}^{\infty} \frac{x^{2k}}{k(2k+1)^2} = -\ln(1-x^2) - \frac{2}{x}\operatorname{Arth} x - \frac{1}{x}[Li_2(x) - Li_2(-x)] + 4,$$

$$\sum_{k=1}^{\infty} \frac{x^{2k}}{k(2k-1)^2} = -\ln(1-x^2) - 2x \operatorname{Arth} x + x[Li_2(x) - Li_2(-x)],$$

$$\sum_{k=1}^{\infty} \frac{x^{2k}}{k^2(2k+1)} = Li_2(x^2) + 2\ln(1-x^2) + \frac{4}{x}\operatorname{Arth} x - 4,$$

$$\sum_{k=1}^{\infty} \frac{x^{2k}}{k^2(2k-1)} = -Li_2(x^2) + 2\ln(1-x^2) + 4x \operatorname{Arth} x,$$

$$\sum_{k=1}^{\infty} \frac{x^{2k}}{k^2(2k+1)^2} = Li_2(x^2) + 4\ln(1-x^2) + \frac{8}{x}\operatorname{Arth} x + \frac{2}{x}[Li_2(x) - Li_2(-x)] - 12,$$

$$\sum_{k=1}^{\infty} \frac{x^{2k}}{k^2(2k-1)^2} = Li_2(x^2) - 4\ln(1-x^2) - 8x \operatorname{Arth} x + 2x[Li_2(x) - Li_2(-x)],$$

and [9, p. 6]

$$Li_2(x^2) = 2[Li_2(x) + Li_2(-x)], \tag{4.24}$$

we can express (4.21) in closed form for $n = 1, 2$ and $m = 1, 2$. After some calculation with REDUCE we obtain for $n = m = 1$:

$$\begin{aligned} &\int_0^{\infty} x^{-3/2} \ln(1+\sigma x) \ln(1+\omega x)\, dx \\ &\quad = 4\pi\left\{\sqrt{\sigma} \ln\left(1+\sqrt{\frac{\omega}{\sigma}}\right) + \sqrt{\omega}\left[\ln\left(1+\sqrt{\frac{\omega}{\sigma}}\right) - \ln\sqrt{\frac{\omega}{\sigma}}\right]\right\}, \end{aligned} \tag{4.25}$$

where, because of the symmetry of the integral with respect to σ and ω, we can assume $|\sigma/\omega| > 1$.

For $n = 1$, $m = 2$, formula (4.21) reduces to

(4.26)

$$\begin{aligned} &\int_0^{\infty} x^{-3/2} \ln(1+\sigma x) Li_2(-\omega x)\, dx \\ &\quad = 4\pi\left\{2\sqrt{\sigma}\left[Li_2\left(-\sqrt{\frac{\omega}{\sigma}}\right) - \ln\left(1+\sqrt{\frac{\omega}{\sigma}}\right)\right] - \sqrt{\omega}\left[2\ln\left(1+\sqrt{\frac{\omega}{\sigma}}\right) - \ln\frac{\omega}{\sigma}\right]\right\} \\ &\qquad (|\sigma/\omega| > 1), \\ &\quad = -\pi\left\{8\sqrt{\omega}\ln\left(1+\sqrt{\frac{\sigma}{\omega}}\right)\right. \\ &\qquad \left. + \sqrt{\sigma}\left[8\ln\left(1+\sqrt{\frac{\sigma}{\omega}}\right) + \ln^2\frac{\sigma}{\omega} - 4\ln\frac{\sigma}{\omega} + 8Li_2\left(-\sqrt{\frac{\sigma}{\omega}}\right) + \frac{4}{3}\pi^2\right]\right\} \\ &\qquad (|\sigma/\omega| < 1). \end{aligned}$$

The same expression, with σ and ω interchanged, holds for $n = 2$, $m = 1$.

For $n = m = 2$, we obtain from (4.21)

$$\int_0^\infty x^{-3/2} Li_2(-\sigma x) Li_2(-\omega x)\, dx$$

$$= 2\pi \left\{ 8\sqrt{\sigma} \left[2 \ln\left(1+\sqrt{\frac{\omega}{\sigma}}\right) - Li_2\left(-\sqrt{\frac{\omega}{\sigma}}\right) \right] \right. \tag{4.27}$$

$$\left. + \sqrt{\omega} \left[16 \ln\left(1+\sqrt{\frac{\omega}{\sigma}}\right) + \ln^2 \frac{\omega}{\sigma} - 8 \ln \frac{\omega}{\sigma} + 8 Li_2\left(-\sqrt{\frac{\omega}{\sigma}}\right) + \frac{4}{3}\pi^2 \right] \right\},$$

where, as for (4.25), we can assume $|\sigma/\omega| > 1$.

For $\sigma = \omega = 1$, the above results lead to

$$\int_0^\infty x^{-3/2} \ln^2(1+x)\, dx = 8\pi \ln 2,$$

$$\int_0^\infty x^{-3/2} \ln(1+x) Li_2(-x)\, dx = -\frac{2}{3}\pi(\pi^2 + 24 \ln 2), \tag{4.28}$$

$$\int_0^\infty x^{-3/2} [Li_2(-x)]^2\, dx = \frac{8}{3}\pi(\pi^2 + 24 \ln 2).$$

Acknowledgment. K. S. Kölbig would like to thank H. Lipps of CERN for a helpful discussion.

REFERENCES

[1] V. S. ADAMCHIK AND O. I. MARICHEV, *On the representation of functions of hypergeometric type in logarithmic cases*, Vestsī Akad. Navuk BSSR Ser. Fīz.-Mat. Navuk, 5 (1983), pp. 29-35. (In Russian.)

[2] B. C. BERNDT AND P. T. JOSHI, *Chapter 9 of Ramanujan's second notebook*, in Contemporary Mathematics Vol. 23, American Mathematical Society, Providence, RI, 1983.

[3] J. BÖHM AND E. HERTEL, *Polyedergeometrie in n-dimensionalen Räumen konstanter Krümmung*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1980.

[4] A. DEVOTO AND D. W. DUKE, *Table of integrals and formulae for Feynman diagram calculations*, Riv. Nuovo Cimento(3), 7 (1984), pp. 1-39.

[5] R. GASTMANS AND W. TROOST, *On the evaluation of polylogarithmic integrals*, Simon Stevin (Ghent), 55 (1981), pp. 205-219.

[6] I. S. GRADSHTEYN AND I. M. RYZHIK, *Table of Integrals, Series, and Products*, Academic Press, New York, 1980.

[7] A. C. HEARN, REDUCE *User's Manual*, version 3.2, Rand Publ. CP78 (Rev. 4/85), Santa Monica CA, 1985.

[8] K. S. KÖLBIG, *Nielsen's generalized polylogarithms*, SIAM J. Math. Anal., 17 (1986), pp. 1232-1258.

[9] L. LEWIN, *Polylogarithms and Associated Functions*, North-Holland, New York, 1981.

[10] W. MAIER AND H. KIESEWETTER, *Funktionalgleichungen mit analytischen Lösungen*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1971.

[11] O. I. MARICHEV, *A method for the evaluation of integrals with hypergeometric functions*, Dokl. Akad. Nauk BSSR, 25 (1981), pp. 590-593. (In Russian.)

[12] ———, *Handbook of Integral Transforms of Higher Transcendental Functions*, Ellis Horwood, Chichester, 1982.

[13] A. M. MATHAI AND R. K. SAXENA, *Generalized Hypergeometric Functions with Applications in Statistics and Physical Sciences*, Lecture Notes in Mathematics 348, Springer, Berlin, 1973.

[14] ———, *The H-Function with Applications in Statistics and Other Disciplines*, Wiley Eastern, New Delhi, 1978.

[15] C. S. MEIJER, *On the G-function*, Proc. Kon. Ned. Akad. Wet., 49 (1946), pp. 227-237.

[16] A. P. PRUDNIKOV, YU. A. BRYCHKOV, AND O. I. MARICHEV, *Integrals and Series, Vol. 1: Elementary Functions*, Gordon and Breach, New York, 1986.

[17] I. N. SNEDDON, *The Use of Integral Transforms*, McGraw-Hill, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

013

SPECTRAL MEASURES, ORTHOGONAL POLYNOMIALS, AND ABSOLUTE CONTINUITY*

JOANNE DOMBROWSKI†

Abstract. This paper studies the spectral measure of an unbounded tridiagonal matrix operator for which the matrix entries satisfy a certain growth condition, and presents a sufficient condition for the existence of an absolutely continuous part. The results are related to a class of orthogonal polynomials with exponential weights.

Key words. absolute continuity, commutators, orthogonal polynomials

AMS(MOS) subject classification. primary 47B15

1. Introduction. The purpose of this paper is to continue the study of unbounded tridiagonal matrix operators, measures and systems of orthogonal polynomials begun in [2]. A brief review of some known results in the bounded case will introduce the unbounded problem to be considered.

A bounded cyclic self-adjoint operator C, with cyclic vector ϕ, defined on a separable Hilbert space $\mathcal{H}$, can be represented as a tridiagonal matrix with respect to the basis obtained by orthonormalizing $\{C^n\phi\}_{n=0}^{\infty}$. The positive subdiagonal sequence $\{a_n\}$ and diagonal sequence $\{b_n\}$ in this matrix can be used to obtain information about the Borel measure $\mu(\beta)=\|E(\beta)\phi_1\|^2$, obtained from the spectral resolution $C=\int \lambda\, dE_\lambda$. It is shown in [1] and [4], for example, that if $\lim a_n = a$, $a\neq 0$, $\lim b_n = 0$, $\sum |a_n - a_{n-1}|<\infty$ and $\sum |b_n - b_{n-1}|<\infty$ then μ restricted to $(-2a, 2a)$ is absolutely continuous with respect to Lebesgue measure. This result was motivated by and is applicable to the study of orthogonal polynomials. For the spectral measure μ is also the measure of orthogonality for the sequence of polynomials $\{P_n\}$ recursively defined as follows:

$$P_1(\lambda)=1, \qquad P_2(\lambda)=\frac{\lambda-b_1}{a_1},$$
$$P_n(\lambda)=\frac{(\lambda-b_n)P_{n-1}(\lambda)-a_{n-2}P_{n-2}(\lambda)}{a_{n-1}}. \tag{1.1}$$

In fact, the polynomials $\{P_n\}$ form an orthonormal basis for $L^2(\mu)$ and C is unitarily equivalent to the multiplication operator on $L^2(\mu)$ defined by $Mf(\lambda)=\lambda f(\lambda)$. Such systems of polynomials have been studied extensively in the literature. One item of interest, among many, has been the relationship between the recurrence coefficients and the nature of the measure of orthogonality.

Recently there has been considerable interest in the study of systems of the form (1.1) with $a_n>0$, b_n real, for which the support of the measure of orthogonality is an unbounded set. In this case the sequence $\{a_n\}$ is unbounded. As discussed in [2], the corresponding tridiagonal matrix $C=[c_{ij}]$ with $c_{ii}=b_i$ and $c_{i,i+1}=c_{i+1,i}=a_i$ defines an unbounded operator on l^2 with domain consisting of those elements in l^2 for which matrix multiplication yields a vector in l^2. If $\sum (1/a_n)=\infty$ then C is self-adjoint and hence has a spectral decomposition $C=\int \lambda\, dE\lambda$. If $\{\phi_n\}$ is the standard basis for l^2, then ϕ_1 is a cyclic vector for C and $\mu(\beta)=\|E(\beta)\phi_1\|^2$ is the measure of orthogonality for the polynomials in (1.1). The purpose of this paper is to present a sufficient condition

* Received by the editors May 12, 1986; accepted for publication June 8, 1987.
† Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435.

in terms of the sequences $\{a_n\}$ and $\{b_n\}$ for the existence of a nontrivial absolutely continuous part for the measure μ. The condition to be presented seems to be the natural generalization to the unbounded case of the condition given above for the bounded case. Whereas in the bounded case it is sufficient that limits exist and differences are absolutely summable, in the unbounded case it is sufficient that limits exist and differences of differences are absolutely summable. This will be made more precise below. The main result will be illustrated with a class of orthogonal polynomials introduced by G. Freud.

2. Main results. Henceforth C will denote an infinite matrix of the form

$$C=\begin{bmatrix} 0 & a_1 & 0 & 0 & \cdots \\ a_1 & 0 & a_2 & 0 & \cdots \\ 0 & a_2 & 0 & a_3 & \cdots \\ 0 & 0 & a_3 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \end{bmatrix} \tag{2.1}$$

with $a_n>0$ and $\lim a_n=\infty$. It will further be assumed that $\sum(1/a_n)=\infty$ (so that C is self-adjoint), that $\sum_{n=2}^{\infty}[a_n^2-a_{n-1}^2]^-<\infty$ and that $\sum|d_n-d_{n-1}|<\infty$, where $d_n=|a_n-a_{n-1}|$. The domain of C will consist of those vectors in l^2 for which matrix multiplication yields a vector in l^2. As shown below, such operators have no eigenvalues. Therefore the spectrum coincides with the essential spectrum and remains fixed if a finite number of terms in the sequence $\{a_n\}$ are changed. This is needed for the main result.

To establish the results on eigenvalues and the existence of an absolutely continuous part the following notation is needed. If $\{\phi_n\}$ is the standard basis for l^2 then $C\phi_n=\frac{1}{2}(T+T^*)\phi_n$ where $T\phi_n=2a_n\phi_{n+1}$. Let $J\phi_n=(1/2i)(T-T^*)\phi_n$ and obtain the bounded operator J_N from J by substituting a_N for a_n when $n\geqq N$. It follows that $CJ_N-J_NC=-2iK_N$ where $K_N=[k_{ij}]$ is a bounded operator with $k_{ii}=a_i^2-a_{i-1}^2$ for $i=1,\cdots,N$, $k_{ii}=a_N(a_i-a_{i-1})$ for $i>N$, $k_{i,i+2}=k_{i+2,i}=\frac{1}{2}a_N(a_{i+1}-a_i)$ for $i\geqq N$ and all other entries equal to zero. This commutator equation, which holds only on a dense subset of H, is fundamental to the arguments to be presented.

The following result essentially appears in [2]. The proof is summarized to indicate the modifications needed for the general setting of this paper. Note that $d_n=|a_n-a_{n-1}|$. Recall from [1] that an induction argument can be used to show that the polynomials defined in (1.1) satisfy the equation

$$a_1^2P_1^2(\lambda)+\sum_{n=2}^{N}(a_n^2-a_{n-1}^2)P_n^2(\lambda)=\left[a_{N-1}P_{N-1}(\lambda)-\frac{\lambda}{2}P_N(\lambda)\right]^2+\left(a_N^2-\frac{\lambda^2}{4}\right)P_N^2(\lambda).$$

THEOREM 1. *If* $\sum[a_n-a_{n-1}]^-<\infty$ *and* $\sum|d_n-d_{n-1}|<\infty$ *then* C *has no eigenvalues.*

Proof. Assume λ is an eigenvalue. One corresponding eigenvector must be $x=\{P_n(\lambda)\}$. Choose N_0 such that for $n\geqq N_0$, $\sum_n^{\infty}[a_i-a_{i-1}]^-<\frac{1}{8}(a_n-|\lambda|/2)$, $d_n<\frac{1}{4}(a_n-|\lambda|/2)$ and $\sum_n^{\infty}|d_i-d_{i-1}|<\frac{1}{4}(a_n-|\lambda|/2)$. Let N be defined by $P_N^2(\lambda)=\max_{n\geqq N_0}P_n^2(\lambda)$. It then follows that

$$\begin{aligned}\langle K_Nx,x\rangle &\geqq\left[a_{N-1}P_{N-1}(\lambda)-\frac{\lambda}{2}P_N(\lambda)\right]^2+\left(a_N^2-\frac{\lambda^2}{4}\right)P_N^2(\lambda)+a_N\sum_{N+1}^{\infty}d_iP_i^2(\lambda)\\ &\quad-2a_N\sum_{N+1}^{\infty}[a_i-a_{i-1}]^-P_i^2(\lambda)-\frac{1}{2}a_N\sum_{N+1}^{\infty}d_i[P_{i+1}^2(\lambda)+P_{i-1}^2(\lambda)]\\ &\geqq\left[a_{N-1}P_{N-1}(\lambda)-\frac{\lambda}{2}P_N(\lambda)\right]^2+\frac{1}{4}\left(a_N^2-\frac{\lambda^2}{4}\right)P_N^2(\lambda).\end{aligned}$$

But this contradicts the fact, established in [2], that if $\{d_n\}$ is bounded and $Cx = \lambda x$ then $\langle K_N x, x\rangle = 0$. □

If C given by (2.1) is self-adjoint with spectral resolution $C = \int \lambda\, dE_\lambda$ then the polynomials defined in (1.1) are orthonormal with respect to the measure $\mu(\beta) = \|E(\beta)\phi_1\|^2$. The following technical lemma about these polynomials is needed for the result on absolute continuity. Note that the lemma has content when a_1^2 is large relative to the sum $\sum_{n=2}^{\infty}[a_n^2 - a_{n-1}^2]^-$. Obviously, for example, the lemma provides information when $\{a_n\}$ is monotone increasing.

LEMMA 1. *Suppose there exists a subinterval* Δ *of* $[-2a_1, 2a_1]$ *and a* $\delta > 0$ *such that* $\lambda \in \Delta$ *implies that* $4a_1^2 - \lambda^2 \geqq \delta$ *and* $\sum_{n=2}^{\infty}[a_n^2 - a_{n-1}^2]^- < \delta/8$. *Then for* $n > 1$, $\int_\Delta P_n^2\, d\mu \leqq 8a_1^2\mu(\Delta)/\delta$.

Proof. Fix $n > 1$. Choose $N < n$ such that $\int_\Delta P_N^2\, d\mu = \max_{1<i\leqq n}\int_\Delta P_i^2\, d\mu$. Then

$$a_1^2\int_\Delta P_1\, d\mu = a_1^2\int_\Delta P_1\, d\mu + \sum_{i=2}^{N}(a_i^2 - a_{i+1}^2)\int_\Delta P_i^2\, d\mu - \sum_{i=2}^{N}(a_i^2 - a_{i-1}^2)\int_\Delta P_i^2\, d\mu$$

$$\geqq \int_\Delta \left(a_N^2 - \frac{\lambda^2}{4}\right)P_N^2\, d\mu - \sum_{i=2}^{N}(a_i^2 - a_{i-1}^2)^+\int_\Delta P_i^2\, d\mu.$$

Since $a_N^2 - a_1^2 = \sum_{i=1}^{N}[a_i^2 - a_{i-1}^2]^+ - \sum_{i=2}^{N}[a_i^2 - a_{i-1}^2]^-$ it follows that

$$a_1^2\int_\Delta P_1\, d\mu \geqq \int_\Delta \left(a_1^2 - \frac{\lambda^2}{4}\right)P_N^2\, d\mu - \frac{\delta}{8}\int_\Delta P_N^2\, d\mu.$$

Hence $(8a_1^2/\delta)\mu(\Delta) \geqq \int_\Delta P_N^2\, d\mu \geqq \int_\Delta P_n^2\, d\mu$ as was to be shown. □

This lemma will now be used to present the main result of this paper. The notation sp (C) will be used for the spectrum of C.

THEOREM 2. *If* $\sum_{n=2}^{\infty}[a_n^2 - a_{n-1}^2]^- < \infty$ *and* $\sum_{n=2}^{\infty}|d_n - d_{n-1}| < \infty$ *then* μ *has an absolutely continuous part with support* sp (C).

Proof. Fix $R \geqq 1$. By the Kato-Rosenblum Theorem [3], [7] it is enough to show that the spectral measure of the trace class perturbation of C obtained by changing a finite number of weights in $\{a_n\}$ is absolutely continuous on $[-R, R]$. Observe that since there are no eigenvalues, all such perturbations of C will have the same spectrum. Now choose N such that $2a_N - R \geqq 8\sum_{n=N}^{\infty}|d_n - d_{n-1}|$, $2a_N - R \geqq 32\sum_{n=N}^{\infty}[a_n^2 - a_{n-1}^2]^-$ and such that for $n \geqq N$, $a_n \geqq R$ and $a_n \geqq \frac{1}{2}a_N$. Note that if N is so chosen, then for $n \geqq N$, $\Delta \subset [-R_1 R]$ and $\lambda \in \Delta$, it follows that $4a_n^2 - \lambda^2 \geqq 4a_n^2 - R^2 \geqq 16a_n\sum_N^{\infty}|d_n - d_{n-1}|$. Similarly, $4a_n^2 - \lambda^2 \geqq 64a_n\sum_N^{\infty}[a_n^2 - a_{n-1}^2]^-$. Assume, for now, that $a_1 = a_2 = \cdots = a_N$ (i.e., consider a trace class perturbation of the given operator). If Δ is a subinterval of $[-R, R]$, then $E(\Delta)\phi_1 = \sum_{i=1}^{\infty}\langle E(\Delta)\phi_1, \phi_i\rangle\phi_i$ where $\langle E(\Delta)\phi_1, \phi_i\rangle = \int_\Delta P_i\, d\mu$. The commutator equation $CJ_N - J_N C = -2iK_N$ is used in [2] to show that $|\langle K_N E(\Delta)\phi_1, E(\Delta)\phi_1\rangle| \leqq \frac{1}{2}\|J_N\|\,|\Delta|\,\|E(\Delta)\phi_1\|^2$. On the other hand,

$$\langle K_N E(\Delta)\phi_1, E(\Delta)\phi_1\rangle = \sum_1^N (a_i^2 - a_{i-1}^2)\left|\int_\Delta P_i\, d\mu\right|^2 + \sum_{N+1}^{\infty} a_N(a_i - a_{i-1})\left|\int_\Delta P_i\, d\mu\right|^2$$

$$+ \sum_{N+1}^{\infty} a_N(a_i - a_{i-1})\int_\Delta P_{i+1}\, d\mu\int_\Delta P_i\, d\mu$$

$$\geqq a_1^2\left|\int P_1\, d\mu\right|^2 + \sum_{N+1}^{\infty} a_N(a_i - a_{i-1})\left|\int_\Delta P_i\, d\mu\right|^2$$

$$-\frac{1}{2}\sum_{N+1}^{\infty} a_N d_i\left[\left|\int_\Delta P_{i+1}\, d\mu\right|^2 + \left|\int_\Delta P_{i-1}\, d\mu\right|^2\right]$$

$$\geqq a_1^2\left|\int_\Delta P_1\,d\mu\right|^2+a_N\left[(a_{N+1}-a_N)-\frac{1}{2}d_{N+1}\right]\left|\int_\Delta P_{N+1}\,d\mu\right|^2$$

$$+\sum_{N+2}^{\infty}a_N[(a_i-a_{i-1})-d_i]\left|\int_\Delta P_i\,d\mu\right|^2$$

$$-\sum_{N}^{\infty}a_N|d_i-d_{i-1}|\left|\int_\Delta P_i\,d\mu\right|^2$$

$$\geqq a_1^2\left|\int_\Delta P_1\,d\mu\right|^2-2\sum_{N+1}^{\infty}a_N[a_i-a_{i-1}]^-\left|\int_\Delta P_i\,d\mu\right|^2$$

$$-\sum_{N}^{\infty}a_N|d_i-d_{i-1}|\left|\int_\Delta P_i\,d\mu\right|^2$$

$$\geqq a_1^2\left|\int_\Delta P_1\,d\mu\right|^2-2\sum_{N+1}^{\infty}(a_i^2-a_{i-1}^2)^-\left|\int_\Delta P_i d\mu\right|^2$$

$$-\sum_{N}^{\infty}a_N|d_i-d_{i-1}|\left|\int_\Delta P_i\,d\mu\right|^2.$$

Now apply Lemma 1 with $\delta=4a_N^2-R^2$. Recall that $\Delta\subset[-R,R]$. It follows that

$$\langle K_N E(\Delta)\phi_1, E(\Delta)\phi_1\rangle\geqq\frac{a_N^2}{8}|\mu(\Delta)|^2.$$

Combining this result with the inequality obtained from the commutator equation implies that $|\mu(\Delta)|\leqq 4a_N^2\|J_N\||\Delta|$. Let β be a Borel subset of $[-R,R]$ of Lebesgue measure zero. Then for any $\varepsilon>0$ there exists a pairwise disjoint sequence of intervals $\{\Delta_j\}$ such that $\beta\subset\cup\,\Delta_j$ and $\sum|\Delta_j|<\varepsilon$. Since $\mu(\beta)\leqq\sum\mu(\Delta_j)\leqq 4a_N^2\|J_N\|\sum|\Delta_j|$ it follows that $\mu(\beta)=0$. Hence it has been shown that the spectral measure of a trace class perturbation of C is absolutely continuous with respect to Lebesgue measure on $[-R,R]$. The theorem follows from an application of the Kato-Rosenblum Theorem.

3. Examples. In this section examples will be presented to illustrate the above results. The asymptotic expansions needed for these examples are developed in [5]. See also the references cited therein.

The simplest example of practical interest is obtained by letting $a_n=\sqrt{n/2}$. It is easily checked that the conditions of the above theorems are satisfied. The corresponding polynomials $\{P_n\}$ are the Hermite polynomials and it is well known that $\mu(\beta)=\int_\beta e^{-x^2}\,dx$.

For a related class of examples choose α to be an even positive integer and let $\{P_n(\lambda)\}$ be the sequence of polynomials obtained by orthonormalizing the sequence $\{\lambda^n\}_{n=0}^{\infty}$ with respect to the measure $\mu(\beta)=\int_\beta e^{-x^\alpha/\alpha}\,dx$. These polynomials satisfy a recursion formula of the form (1.1) with $b_n=0$ and, as shown in [5],

$$a_n=n^{1/\alpha}\left[c_0+\frac{c_2}{n^2}+O\left(\frac{1}{n^4}\right)\right].$$

Since

$$a_n-a_{n-1}=c_0[n^{1/\alpha}-(n-1)^{1/\alpha}]+c_2[n^{-2+(1/\alpha)}-(n-1)^{-2+(1/\alpha)}]+n^{1/\alpha}t_n-(n-1)^{1/\alpha}t_{n-1}$$

with $|t_n| \leqq M/n^4$ for all n, the Mean Value Theorem applied to $f(x) = x^{1/\alpha}$ shows that for large n, $a_n \geqq a_{n-1}$. That is, the term $c_0[n^{1/\alpha} - (n-1)^{1/\alpha}]$ is positive and, for large n, dominates the remaining terms of the sum. For large n, write $d_n = a_n - a_{n-1} = x_n + y_n$ with $x_n = c_0[n^{1/\alpha} - (n-1)^{1/\alpha}]$. Note that $\{x_n\}$ is decreasing and $\sum |y_n| < \infty$. It follows that $\sum |d_n - d_{n-1}| \leqq \sum |x_n - x_{n-1}| + \sum |y_n - y_{n-1}| < \infty$. Therefore the hypotheses of the above theorems are satisfied.

4. Final comments. The main theorem of this paper presents a sufficient condition for the existence of an absolutely continuous part for the measure μ. It does not, in contrast to the results presented in [2], claim that μ is absolutely continuous. This is, however, true for the examples cited above. It would be interesting to know if the hypotheses of the main theorem are sufficient to conclude that μ is indeed absolutely continuous.

REFERENCES

[1] J. DOMBROWSKI, *Tridiagonal matrix representations of cyclic self-adjoint operators* II, Pacific J. Math., 120 (1985), pp. 47-53.

[2] ———, *Cyclic operators, commutators and absolutely continuous measures*, Proc. Amer. Math. Soc., 100 (1987), pp. 457-463.

[3] T. KATO, *Perturbation of continuous spectra by trace class operators*, Proc. Japan Acad., 33 (1957), pp. 260-264.

[4] A. MATÉ AND P. NEVAI, *Orthogonal polynomials and absolutely continuous measures*, in Approximation Theory, IV, C. K. Chui, L. L. Schumakar, and J. D. Ward, eds. Academic Press, New York, 1983, pp. 611-617.

[5] A. MATÉ, P. NEVAI, AND T. ZASLAVSKY, *Asymptotic expansion of ratios of coefficients of orthogonal polynomials with exponential weights*, Trans. Amer. Math. Soc., 287 (1985), pp. 495-505.

[6] C. R. PUTNAM, *Commutation Properties of Hilbert Space Operators and Related Topics*, Ergebnesse der Math. 36, Springer, Berlin, 1967.

[7] M. ROSENBLUM, *Perturbation of the continuous spectrum and unitary equivalence*, Pacific J. Math., 7 (1957), pp. 997-1010.

[8] M. STONE, *Linear Transformations in Hilbert Space*, American Mathematical Society, New York, 1932.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

014

A PROOF OF SOME q-ANALOGUES OF SELBERG'S INTEGRAL FOR $k=1$*

KEVIN W. J. KADELL†

Abstract. Selberg has given an important multiple beta type integral. We conjecture that for all $k \geqq 0$, there exists a family $\{s^k_{n,\lambda}(\mathbf{t})\}$ of homogeneous symmetric polynomials with the following property. If the integrand in Selberg's integral is multiplied by $s^k_{n,\lambda}(\mathbf{t})$, then the integral still has a simple closed form. For all $k \geqq 0$, this family should include the elementary symmetric functions.

For $k=1$, our symmetric functions are the Schur functions. We prove some q-analogues of this and some related results.

Key words. Selberg's integral, Schur functions, Selberg polynomials, q-beta integral

AMS(MOS) subject classification. 33A15

1. Introduction and summary. Selberg [27] has given an important multiple beta type integral. It is the case $m=l=0$ of Conjecture 1.

Conjecture 1.

$$I_{n,m,l}(x,y,k)=\int_0^1\cdots\int_0^1\prod_{i=1}^n t_i^{(x-1)+\chi(i\leqq m)}(1-t_i)^{(y-1)+\chi(n-i+1\leqq l)}\Delta_n^{2k}(\mathbf{t})\,dt_1\cdots dt_n$$

$$=\prod_{i=1}^n\frac{\Gamma(x+(n-i)k+\chi(i\leqq m))\Gamma(y+(n-i)k+\chi(i\leqq l))\Gamma(1+ik)}{\Gamma(x+y+(2n-i-1)k+\chi(i\leqq m+l))\Gamma(1+k)}, \tag{1.1}$$

where $\chi(A)$ is 1 or 0 according to whether A is true or false,

$$\Delta_n(t_1,\cdots,t_n)=\Delta_n(\mathbf{t})=\prod_{1\leqq i<j\leqq n}(t_i-t_j), \tag{1.2}$$

and, as holds throughout, $\operatorname{Re}(x)>0$, $\operatorname{Re}(y)>0$ and n, m, l, and k are nonnegative integers satisfying $m+l\leqq n$. We omit l when $l=0$. The substitution $t_i\to(1-t_{n-i+1})$, $1\leqq i\leqq n$, gives the symmetry

$$I_{n,m,l}(x,y,k)=I_{n,l,m}(y,x,k). \tag{1.3}$$

We can probably extend Selberg's proof to treat Conjecture 1 using (1.3).

Bombieri and Selberg recently observed that Selberg's integral includes a conjecture of Mehta [22, p. 42] as a limiting case. Macdonald [20] and Morris [24] used it to establish the case $q=1$ of some constant term conjectures associated with certain root systems. See Evans [11] for character sum analogues of these results. Askey [6] gives a number of conjectured q-analogues of Selberg's integral and relates them to the case $a_1=a_2=\cdots=a_n$ of Andrews' q-Dyson conjecture [1]. Zeilberger and Bressoud [29] have recently proven this conjecture.

We have the following conjecture.

Conjecture 2. Let λ denote a partition $\lambda_1\geqq\lambda_2\geqq\cdots\geqq\lambda_n\geqq 0$ with at most n parts. For all $k\geqq 0$ there exists a homogeneous symmetric polynomial $s^k_{n,\lambda}(\mathbf{t})$ with leading

* Received by the editors August 12, 1985; accepted for publication May 28, 1987. Most of this work was performed during the author's postdoctoral appointment at the University of Wisconsin, Madison, Wisconsin 53706, supported by a National Science Foundation grant. This work was partially supported by a faculty grant-in-aid from Arizona State University.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287. Present address, School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

term $\prod_{i=1}^{n} t_i^{\lambda_i}$ such that

$$
\begin{aligned}
I_{n,\lambda}(x, y, k) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)}(1-t_i)^{(y-1)} s_{n,\lambda}^{k}(\mathbf{t}) \Delta_n^{2k}(\mathbf{t})\, dt_1 \cdots dt_n \\
&= n!\, f_{n,\lambda}^{k} \prod_{i=1}^{n} \frac{\Gamma(x+(n-i)k+\lambda_i)\Gamma(y+(n-i)k)}{\Gamma(x+y+(2n-i-1)k+\lambda_i)},
\end{aligned} \tag{1.4}
$$

where

$$
f_{n,\lambda}^{k} = \prod_{1 \le i < j \le n} \prod_{u=0}^{k-1} (k(j-i)+\lambda_i-\lambda_j+u). \tag{1.5}
$$

We call these the Selberg polynomials. Mena [23] and Richards [26] obtained a similar set of polynomials. They used a different basis which retains the symmetry in x and y.

Clearly $s_{n,\lambda}^{0}(\mathbf{t})$ are the monomial symmetric functions. We shall show that

$$
s_{n,\lambda}^{1}(\mathbf{t}) = \frac{\det |t_j^{n-i+\lambda_i}|_{n\times n}}{\det |t_j^{n-i}|_{n\times n}} \tag{1.6}
$$

are the Schur symmetric functions. Conjecture 1 is motivated by the fact that the elementary symmetric functions are both monomial and Schur functions. Equation (1.6) includes the case $l=0$ of Conjecture 1 when $k=1$. The restriction on l is easily removed. We prove q-analogues of all of these results for $k=1$.

Fix q with $0<q<1$ and set

$$
\begin{aligned}
&(x)_0 = (x; q)_0 = 1, \\
&(x)_n = (x; q)_n = \prod_{i=0}^{n-1} (1-xq^i), \qquad n \ge 1, \\
&(x)_\infty = (x; q)_\infty = \lim_{n\to\infty} (x)_n = \prod_{i=0}^{\infty} (1-xq^i).
\end{aligned} \tag{1.7}
$$

We usually omit the base q. Following Jackson [13] we set

$$
\int_0^1 f(t)\, d_q t = (1-q) \sum_{n=0}^{\infty} q^n f(q^n) \tag{1.8}
$$

and

$$
\Gamma_q(x) = (1-q)^{(1-x)} \frac{(q)_\infty}{(q^x)_\infty}. \tag{1.9}
$$

Askey's first conjectured q-analogue [6] of Selberg's integral is based upon the q-analogue

$$
\int_0^1 t^{(x-1)} \frac{(tq)_\infty}{(tq^y)_\infty}\, d_q t = \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)} \tag{1.10}
$$

of the beta integral. See Askey [5]. It is

$$
\begin{aligned}
&\int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)} \frac{(t_i q)_\infty}{(t_i q^y)_\infty} \prod_{1\le i<j\le n} t_i^{2k} \left(\frac{t_j q^{1-k}}{t_i}\right)_{2k} d_q t_1 \cdots d_q t_n \\
&\quad = q^{[kx\binom{n}{2}+2k^2\binom{n}{3}]} \prod_{i=1}^{n} \frac{\Gamma_q(x+(n-i)k)\Gamma_q(y+(n-i)k)\Gamma_q(1+ik)}{\Gamma_q(x+y+(2n-i-1)k)\Gamma_q(1+k)}.
\end{aligned} \tag{1.11}
$$

Selberg's proof could not be used directly on (1.11), since the exponent of q is not symmetric in x and y. Macdonald [19] has proven (1.11) when $k=1$ by using a basic property [18, Chap. I, (5.11)] of the skew Schur symmetric functions.

In § 2, we give a recurrence relation for $I_{n,m}(x, y, k)$ which holds for all $k \geqq 0$. In § 3, we prove Conjecture 1 for $k=1$ and treat (1.6). This gives our basic proof of the case $k=1$.

In § 4, we give some preliminary results which we require for the q-case. We reformulate (1.11) with an integrand that is symmetric in $t_1, t_2, \cdots, t_n$ and generalize the Vandermonde determinant.

In § 5, we insert the parameters m and l into Askey's conjecture (1.11). We establish the case $n=2$, $l=0$. This result holds for all $k \geqq 0$. We obtain the sum of a nearly poised ${}_3\phi_2$ which is a companion to Carlitz's q-analogue [9] of Dixon's theorem (Bailey [8, § 3.1, (1)]).

In § 6, we prove our q-analogue of Conjecture 1 for $k=1$, $l=0$. In § 7, we prove a q-analogue of Conjecture 2 for $k=1$. In § 8, we extend the analysis to treat $l>0$ for any $k \geqq 0$. In § 9, we obtain further results by using different q-analogues of the beta integral. Surprisingly, each of these cases leads to the same determinant, which can be evaluated by Lemma 5 of § 4.

2. A recurrence relation. We obtain a recurrence relation which holds for all $k \geqq 0$. Observe that

$$I_{n,m}(x, y, k) = m \int_0^1 t_1^x (1-t_1)^{(y-1)} \left[\int_{t_1}^1 \cdots \int_{t_1}^1 \prod_{i=2}^{n} t_i^{(x-1)+\chi(i \leqq m)} (1-t_i)^{(y-1)} \cdot \Delta_n^{2k}(\mathbf{t})\, dt_2 \cdots dt_n \right] dt_1 \tag{2.1a}$$

$$+ (n-m) \int_0^1 t_n^{(x-1)} (1-t_n)^{(y-1)} \left[\int_{t_n}^1 \cdots \int_{t_n}^1 \prod_{i=1}^{n-1} t_i^{(x-1)+\chi(i \leqq m)} (1-t_i)^{(y-1)} \cdot \Delta_n^{2k}(\mathbf{t})\, dt_1 \cdots dt_{n-1} \right] dt_n. \tag{2.1b}$$

Equation (2.1a) gives the contribution to the integral for $1 \leqq i \leqq m$ where $t_i = \min(t_1, t_2, \cdots, t_n)$ and (2.1b) gives the contribution for $m < i \leqq n$. Let $\operatorname{Re}(y) > 1$. Then

$$\lim_{x \to 0} x \int_0^1 t_1^x (1-t_1)^{(y-1)} \left[\int_{t_1}^1 \cdots \int_{t_1}^1 \prod_{i=2}^{n} t_i^{(x-1)+\chi(i \leqq m)} (1-t_i)^{(y-1)} \cdot \Delta_n^{2k}(\mathbf{t})\, dt_2 \cdots dt_n \right] dt_1$$

$$\leqq \lim_{x \to 0} x \int_0^1 t^x \left[\int_t^1 s^{(x-1)}\, ds \right]^{(n-1)} dt \tag{2.2}$$

$$= \lim_{x \to 0} x^{(2-n)} \int_0^1 t^x (1-t^x)^{(n-1)}\, dt$$

$$= \lim_{x \to 0} x^{(1-n)} \int_0^1 u^{(1/x)} (1-u)^{(n-1)}\, du$$

$$= \lim_{x \to 0} \frac{x\Gamma(n)}{(1+x) \cdots (1+nx)} = 0.$$

Although the restriction $\operatorname{Re}(y) > 1$ can be removed by a careful treatment of the contribution near $t_i = 1$, $1 \leqq i \leqq n$, we prefer to remove it later. We have the following lemma.

LEMMA 3. *Let*

$$f(x, t)=f(x, 0)+O(t) \tag{2.3}$$

with a constant independent of x for some neighborhood of $x=0$ and let $f(x, 0)$ be continuous at $x=0$. Then

$$\lim_{x\to 0} x\int_0^1 t^{(x-1)}(1-t)^{(y-1)}f(x, t)\,dt=f(0, 0). \tag{2.4}$$

Proof. Substituting (2.3) into (2.4), we obtain

$$\begin{aligned}&\lim_{x\to 0} x\int_0^1 t^{(x-1)}(1-t)^{(y-1)}f(x, t)\,dt\\ &\quad=\lim_{x\to 0} x\int_0^1 t^{(x-1)}(1-t)^{(y-1)}(f(x, 0)+O(t))\,dt\\ &\quad=\lim_{x\to 0} x\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}f(x, 0)+O\left(\lim_{x\to 0} x\frac{\Gamma(x+1)\Gamma(y)}{\Gamma(x+y+1)}\right)\\ &\quad=f(0, 0),\end{aligned} \tag{2.5}$$

as required. □

Observe that

$$f(x, t_n)=\int_{t_n}^1\cdots\int_{t_n}^1\prod_{i=1}^{n-1} t_i^{(x-1)+\chi(i\leqq m)}(1-t_i)^{(y-1)}\Delta_n^{2k}(\mathbf{t})\,dt_1\cdots dt_{n-1} \tag{2.6}$$

satisfies our hypotheses. Since

$$\Delta_n^{2k}(\mathbf{t})=\prod_{i=1}^{n-1}(t_i-t_n)^{2k}\Delta_{n-1}^{2k}(\mathbf{t}), \tag{2.7}$$

we have

$$f(0, 0)=I_{n-1,m}(2k, y, k). \tag{2.8}$$

When $m=n$, we must use the convention

$$I_{n-1,n}(x, y, k)=I_{n-1,n-1}(x, y, k). \tag{2.9}$$

Using (2.1a) and (2.1b), we obtain the recurrence relation

$$\lim_{x\to 0} xI_{n,m}(x, y, k)=(n-m)I_{n-1,m}(2k, y, k). \tag{2.10}$$

By (2.2), the contribution of (2.1a) to the limit in (2.10) is 0. The result follows by applying Lemma 3 to (2.1b). This idea is due to Askey.

Let $P_n(\mathbf{t})$ be a symmetric polynomial. Then

$$\begin{aligned}&\lim_{x\to 0} x\int_0^1\cdots\int_0^1\prod_{i=1}^{n} t_i^{(x-1)}(1-t_i)^{(y-1)}P_n(\mathbf{t})\Delta_n^{2k}(\mathbf{t})\,dt_1\cdots dt_n\\ &\quad=n\int_0^1\cdots\int_0^1\prod_{i=1}^{n-1} t_i^{(2k-1)}(1-t_i)^{(y-1)}P_n(t_1, t_2, \cdots, t_{n-1}, 0)\\ &\qquad\cdot\Delta_{n-1}^{2k}(\mathbf{t})\,dt_1\cdots dt_{n-1}.\end{aligned} \tag{2.11}$$

Since $P_n(\mathbf{t})$ is a symmetric function, the integral on the left side of (2.11) is n times the integral taken over the region $t_n=\min(t_1, t_2, \cdots, t_n)$. As in (2.2), we find that any

term ω of $P_n(\mathbf{t})$ which contains t_n to a positive power contributes 0 to the limit in (2.11). Setting $t_n = 0$ we delete these terms from $P_n(\mathbf{t})$. Applying Lemma 3 gives (2.11).

We conjecture that

$$s^k_{n,\lambda}(\mathbf{t}) = \left(\prod_{i=1}^{n} t_i\right)^{\lambda_n} s^k_{n,(\lambda_1-\lambda_n,\cdots,\lambda_{n-1}-\lambda_n,0)}(\mathbf{t}) \tag{2.12}$$

and

$$s^k_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(t_1, t_2, \cdots, t_{n-1}, 0) = s^k_{n-1,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1})}(\mathbf{t}). \tag{2.13}$$

These are easy to see for the Schur functions.

Equation (2.12) gives

$$I_{n,\lambda}(x, y, k) = I_{n,(\lambda_1-\lambda_n,\cdots,\lambda_{n-1}-\lambda_n,0)}(x+\lambda_n, y, k). \tag{2.14}$$

Since $f^k_{n,\lambda}$ is a function of the differences $\lambda_i - \lambda_j$, $1 \leqq i < j \leqq n$, the product on the right side of (1.4) also satisfies (2.14). Thus the parameter λ_n in Conjecture 2 is subsumed by x. Using (2.11), we obtain

$$\lim_{x\to 0} xI_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(x, y, k) = nI_{n-1,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1})}(2k, y, k). \tag{2.15}$$

If $\lambda_n > 0$, then $\lim_{x\to 0} xI_{n,\lambda}(x, y, k)$ is 0.

3. A proof of the case $k = 1$. We have the Vandermonde determinant

$$\begin{aligned}\Delta_n(\mathbf{t}) &= \prod_{1\leqq i<j\leqq n} (t_i - t_j) = \det |t_j^{n-i}|_{n\times n} \\ &= \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} t_i^{(n-\pi(i))}\end{aligned} \tag{3.1}$$

and the expansion

$$\begin{aligned}\Delta_n(\mathbf{t}) &= \prod_{1\leqq i<j\leqq n} [(1-t_j)-(1-t_i)] \\ &= \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} (1-t_i)^{(\pi(i)-1)}\end{aligned} \tag{3.2}$$

about $t_i = 1$, $1 \leqq i \leqq n$.

Let

$$S_{n,m} = \{\pi \in S_n \mid \pi(i) \leqq m \text{ whenever } 1 \leqq i \leqq m\}. \tag{3.3}$$

Clearly $S_{n,0} = S_{n,n} = S_n$. Expanding one of the $2k$ Vandermondes $\Delta_n(\mathbf{t})$ by (3.1), we obtain

$$\begin{aligned}I_{n,m}(x, y, k) = \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i\leqq m)+n-\pi(i)}(1-t_i)^{(y-1)} \\ \cdot \Delta_n^{(2k-1)}(\mathbf{t})\, dt_1 \cdots dt_n.\end{aligned} \tag{3.4}$$

For $\pi \notin S_{n,m}$, it is easy to see that at least two of the exponents

$$(x-1)+n-\pi(i)+\chi(i\leqq m), \qquad 1 \leqq i \leqq n, \tag{3.5}$$

must be equal. Since $\Delta_n^{(2k-1)}(\mathbf{t})$ is an antisymmetric function, the contribution to (3.4) must be 0. Hence, the sum in (3.4) may be restricted to $S_{n,m}$. The Jacobian of the

transformation $t_i \to t_{\pi(i)}$, $1 \leqq i \leqq n$, is sgn (π). Since $\Delta_n^{(2k-1)}(\mathbf{t})$ is an antisymmetric function, each $\pi \in S_{n,m}$ gives the same contribution to the sum in (3.4). Hence

$$I_{n,m}(x, y, k) = m!(n-m)! \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i\leqq m)+(n-i)}(1-t_i)^{(y-1)} \cdot \Delta_n^{(2k-1)}(\mathbf{t})\, dt_1 \cdots dt_n. \tag{3.6}$$

Let (1^m) denote the partition in which the part 1 occurs m times. The elementary symmetric function is given by

$$e_{n,m}(\mathbf{t}) = \sum_{\substack{M\subset\{1,\cdots,n\} \\ |M|=m}} \prod_{i\in M} t_i. \tag{3.7}$$

Our argument shows that

$$\Delta_n(\mathbf{t}) e_{n,m}(\mathbf{t}) = \det |t_j^{(n-i)+\chi(i\leqq m)}|_{n\times n}. \tag{3.8}$$

Hence, by (1.6)

$$s^1_{n,(1^m)}(\mathbf{t}) = e_{n,m}(\mathbf{t}). \tag{3.9}$$

For $k=1$, we expand the Vandermonde in (3.6) using (3.2). This yields

$$\begin{aligned} I_{n,m}(x, y, 1) &= m!(n-m)! \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i\leqq m)+(n-i)} \\ &\quad \cdot (1-t_i)^{(y-1)+(\pi(i)-1)}\, dt_1 \cdots dt_n \\ &= m!(n-m)! \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} \frac{\Gamma(x+n-i+\chi(i\leqq m))\Gamma(y+\pi(i)-1)}{\Gamma(x+y+n-i+\pi(i)-1+\chi(i\leqq m))} \\ &= m!(n-m)! \prod_{i=1}^{n} \Gamma(x+n-i+\chi(i\leqq m))\Gamma(y+n-i)\, U_{n,m}(x, y, 1), \end{aligned} \tag{3.10}$$

where

$$U_{n,m}(x, y, 1) = \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} \frac{1}{\Gamma(x+y+n-i+\pi(i)-1+\chi(i\leqq m))}. \tag{3.11}$$

We wish to prove that Conjecture 1 holds for $l=0$, $k=1$. This is

$$I_{n,m}(x, y, 1) = \prod_{i=1}^{n} \frac{\Gamma(x+n-i+\chi(i\leqq m))\Gamma(y+n-i)}{\Gamma(x+y+2n-i-1+\chi(i\leqq m))}\Gamma(1+i). \tag{3.12}$$

We have explicitly obtained part of this result. Selberg's proof [29] extracts a different set of factors. Observe that in agreement with (3.12), $U_{n,m}(x, y, 1)$ is a function of $x+y$ rather than of x and y. Thus

$$U_{n,m}(x, y, 1) = U_{n,m}(0, x+y, 1). \tag{3.13}$$

To evaluate $U_{n,m}(x, y, 1)$, we proceed by induction on n. Let $0 \leqq m < n$. Since $\Gamma(x)$ occurs as a factor on the right side of (3.10), we obtain

$$\begin{aligned} &\lim_{x\to 0} xI_{n,m}(x, y, 1) \\ &\quad = m!(n-m)!\Gamma(y) \prod_{i=1}^{n-1} \Gamma(n-i+\chi(i\leqq m))\Gamma(y+n-i)\, U_{n,m}(0, y, 1). \end{aligned} \tag{3.14}$$

However, our recurrence relation (2.10) and our induction hypothesis give

$$\lim_{x\to 0} xI_{n,m}(x, y, 1) = (n-m)I_{n-1,m}(2, y, 1)$$

$$= (n-m)\prod_{i=1}^{n-1} \frac{\Gamma(n-i+1+\chi(i\leqq m))\Gamma(y+n-i-1)}{\Gamma(y+2n-i-1+\chi(i\leqq m))}\Gamma(1+i). \tag{3.15}$$

Solving (3.14) and (3.15) for $U_{n,m}(0, y, 1)$, we obtain

$$U_{n,m}(0, y, 1) = \frac{1}{m!(n-m)!}\frac{(n-m)}{\Gamma(y)}\prod_{i=1}^{n-1}\frac{(n-i+\chi(i\leqq m))}{(y+n-i-1)}\frac{\Gamma(1+i)}{\Gamma(y+2n-i-1+\chi(i\leqq m))}$$

$$= \frac{1}{m!(n-m)!}\frac{n!}{\Gamma(y+n-1)}\prod_{i=1}^{n-1}\frac{\Gamma(1+i)}{\Gamma(y+2n-i-1+\chi(i\leqq m))} \tag{3.16}$$

$$= \frac{1}{m!(n-m)!}\prod_{i=1}^{n}\frac{\Gamma(1+i)}{\Gamma(y+2n-i-1+\chi(i\leqq m))}.$$

Replacing y by $x+y$ in (3.16) and using (3.13) yields

$$U_{n,m}(x, y, 1) = \frac{1}{m!(n-m)!}\prod_{i=1}^{n}\frac{\Gamma(1+i)}{\Gamma(x+y+2n-i-1+\chi(i\leqq m))}. \tag{3.17}$$

Substituting (3.17) into (3.10), we get the required result (3.12). The case $m=n$ is equivalent to the case $m=0$ with x replaced by $x+1$. This completes our induction subject to the condition $\mathrm{Re}\,(y)>1$ imposed for (2.2). We may analytically continue both sides of (3.12) to $\mathrm{Re}\,(y)>0$.

Since

$$1 = (1-t_{m+1}) + t_{m+1}, \tag{3.18}$$

we obtain

$$I_{n,m,l}(x, y, k) = I_{n,m,l+1}(x, y, k) + I_{n,m+1,l}(x, y, k). \tag{3.19}$$

Equation (3.12) gives Conjecture 1 for $l=0$, $k=1$. Conjecture 1 follows for $k=1$ by induction on l using (3.19).

We turn to the case $k=1$ of Conjecture 2. This is

$$I_{n,\lambda}(x, y, 1) = \int_0^1\cdots\int_0^1 \prod_{i=1}^{n} t_i^{(x-1)}(1-t_i)^{(y-1)}s^1_{n,\lambda}(\mathbf{t})\Delta_n^2(\mathbf{t})\,dt_1\cdots dt_n$$

$$= n!\prod_{1\leqq i<j\leqq n}(j-i+\lambda_i-\lambda_j)\prod_{i=1}^{n}\frac{\Gamma(x+n-i+\lambda_i)\Gamma(y+n-i)}{\Gamma(x+y+2n-i-1+\lambda_i)}, \tag{3.20}$$

where $s^1_{n,\lambda}(\mathbf{t})$ is the Schur function given by (1.6). Multiplying (1.6) by $\Delta_n(\mathbf{t})$ gives

$$s^1_{n,\lambda}(\mathbf{t})\Delta_n(\mathbf{t}) = \det|t_j^{n-i+\lambda_i}|_{n\times n}. \tag{3.21}$$

Substituting (3.21) into (3.20) yields

$$I_{n,\lambda}(x, y, 1) = \sum_{\pi\in S_n}\mathrm{sgn}\,(\pi)\int_0^1\cdots\int_0^1\prod_{i=1}^{n} t_i^{(x-1)+(n-\pi(i)+\lambda_{\pi(i)})}(1-t_i)^{(y-1)}\Delta_n(\mathbf{t})\,dt_1\cdots dt_n$$

$$= n!\int_0^1\cdots\int_0^1\prod_{i=1}^{n} t_i^{(x-1)+(n-i+\lambda_i)}(1-t_i)^{(y-1)}\Delta_n(\mathbf{t})\,dt_1\cdots dt_n. \tag{3.22}$$

Using (3.2) to expand the Vandermonde in (3.22) yields

$$
\begin{aligned}
I_{n,\lambda}(x,y,1) &= n!\sum_{\pi\in S_n}\operatorname{sgn}(\pi)\int_0^1\cdots\int_0^1\prod_{i=1}^n t_i^{(x-1)+(n-i+\lambda_i)}\\
&\quad\cdot(1-t_i)^{(y-1)+(\pi(i)-1)}\,dt_1\cdots dt_n\\
&= n!\sum_{\pi\in S_n}\operatorname{sgn}(\pi)\prod_{i=1}^n\frac{\Gamma(x+n-i+\lambda_i)\Gamma(y+\pi(i)-1)}{\Gamma(x+y+n-i+\lambda_i+\pi(i)-1)}\\
&= n!\prod_{i=1}^n\Gamma(x+n-i+\lambda_i)\Gamma(y+n-i)\,U_{n,\lambda}(x,y,1),
\end{aligned}
\tag{3.23}
$$

where

$$
U_{n,\lambda}(x,y,1)=\sum_{\pi\in S_n}\operatorname{sgn}(\pi)\prod_{i=1}^n\frac{1}{\Gamma(x+y+n-i+\lambda_i+\pi(i)-1)}.
\tag{3.24}
$$

Observe that

$$
U_{n,\lambda}(x,y,1)=U_{n,(\lambda_1-\lambda_n,\cdots,\lambda_{n-1}-\lambda_n,0)}(0,x+y+\lambda_n,1).
\tag{3.25}
$$

We proceed by induction on n. We treat first the case $\lambda_n=0$. By (3.23) we have

$$
\begin{aligned}
&\lim_{x\to 0} xI_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(x,y,1)\\
&\quad= n!\Gamma(y)\prod_{i=1}^{n-1}\Gamma(n-i+\lambda_i)\Gamma(y+n-i)\,U_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(0,y,1).
\end{aligned}
\tag{3.26}
$$

Our recurrence relation (2.15) and our induction hypothesis give

$$
\begin{aligned}
&\lim_{x\to 0} xI_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(x,y,1)\\
&\quad= n(n-1)!f^1_{n-1,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1})}\prod_{i=1}^{n-1}\frac{\Gamma(n-i+1+\lambda_i)\Gamma(y+n-i-1)}{\Gamma(y+2n-i-1+\lambda_i)}\\
&\quad= n!\prod_{1\le i<j\le n-1}(j-i+\lambda_i-\lambda_j)\prod_{i=1}^{n-1}\frac{\Gamma(n-i+1+\lambda_i)\Gamma(y+n-i-1)}{\Gamma(y+2n-i-1+\lambda_i)}.
\end{aligned}
\tag{3.27}
$$

Solving (3.26) and (3.27) for $U_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(0,y,1)$ yields

$$
\begin{aligned}
&U_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(0,y,1)\\
&\quad=\prod_{1\le i<j\le n-1}(j-i+\lambda_i-\lambda_j)\prod_{i=1}^{n-1}(n-i+\lambda_i)\prod_{i=1}^{n}\frac{1}{\Gamma(y+2n-i-1+\lambda_i)}.
\end{aligned}
\tag{3.28}
$$

Replacing y by $x+y+\lambda_n$ and λ_i, $1\le i\le n-1$, by $\lambda_i-\lambda_n$, $1\le i\le n-1$, in (3.28) and using (3.25) yields

$$
U_{n,\lambda}(x,y,1)=\prod_{1\le i<j\le n}(j-i+\lambda_i-\lambda_j)\prod_{i=1}^{n}\frac{1}{\Gamma(x+y+2n-i-1+\lambda_i)}.
\tag{3.29}
$$

Substituting (3.29) into (3.23), we get the required result (3.20).

4. Preliminaries. Let $\mathbf{Q}=(Q_{i,j})_{n\times n}$ be an upper triangular matrix. Expanding the $\mathbf{Q}$-Vandermonde

$$
{}_{\mathbf{Q}}\Delta_n(\mathbf{t})=\prod_{1\le i<j\le n}(t_i-Q_{i,j}t_j)
\tag{4.1}
$$

in powers of $t_1, t_2, \cdots, t_n$ yields

$$(4.2)\qquad {}_{\mathbf{Q}}\Delta_n(\mathbf{t}) = \sum_{\pi\in S_n} \operatorname{sgn}(\pi)\left(\prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j}\right)\prod_{i=1}^{n} t_i^{(n-\pi(i))} + {}_{\mathbf{Q}}\Sigma_n^{**}(\mathbf{t}),$$

where ${}_{\mathbf{Q}}\Sigma_n^{**}(\mathbf{t})$ is the sum of all of the terms in the expansion of ${}_{\mathbf{Q}}\Delta_n(\mathbf{t})$ in which at least two of the variables $t_1, t_2, \cdots, t_n$ occur to the same power. ${}_{\mathbf{Q}}\Sigma_n^{**}(\mathbf{t})$ expresses the extent to which ${}_{\mathbf{Q}}\Delta_n(\mathbf{t})$ fails to be an antisymmetric function like $\Delta_n(\mathbf{t})$. We define the antisymmetrization of $f_n(\mathbf{t})$ by

$$(4.3)\qquad \begin{aligned} &\Xi_n(f_n(\mathbf{t})) = \sum_{\pi\in S_n} \operatorname{sgn}(\pi) f_n(\pi(\mathbf{t})),\\ &\pi(\mathbf{t}) = (t_{\pi(1)}, t_{\pi(2)}, \cdots, t_{\pi(n)}). \end{aligned}$$

We have the following lemma.

LEMMA 4.

$$(4.4)\qquad \Xi_n\left(\prod_{i=1}^{m} t_i\, {}_{\mathbf{Q}}\Delta_n(\mathbf{t})\right) = \left[\sum_{\pi\in S_{n,m}} \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j}\right] e_{n,m}(\mathbf{t})\Delta_n(\mathbf{t}).$$

Proof. Let ω be a term of ${}_{\mathbf{Q}}\Delta_n(\mathbf{t})$. If t_i and t_j where $1\leqq i<j\leqq n$ occur to the same power in $\omega\prod_{i=1}^{m} t_i$, then, since the left side of (4.4) is an antisymmetric function, we see that the contribution of ω to (4.4) is 0. Assume that all of the exponents in $\omega\prod_{i=1}^{m} t_i$ are distinct. We claim that

$$(4.5)\qquad \omega\prod_{i=1}^{m} t_i = \operatorname{sgn}(\pi)\left(\prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j}\right)\prod_{i=1}^{n} t_i^{(n-\pi(i))+\chi(i\leqq m)},$$

where

$$(4.6)\qquad \pi\in S_{n,m}.$$

This is clear for $m=0$. Assume that $m>0$. Since the maximum exponent in $\omega\prod_{i=1}^{m} t_i$ is n and they are distinct, there is a single possibility which fails to occur. The total of all of the exponents of $\omega\prod_{i=1}^{m} t_i$ is $m+\binom{n}{2}$, so the missing exponent is $n-m$. Since $m>0$, there exists k, $1\leqq k\leqq n$, such that t_k occurs to the power n in $\omega\prod_{i=1}^{m} t_i$. The exponent of t_k in ω is at most $n-1$. Hence $1\leqq k\leqq m$ and t_k is chosen from each possible factor of ${}_{\mathbf{Q}}\Delta_n(\mathbf{t})$. We may take $\pi(k)=1$ and our claim (4.6) follows by induction with n, m, replaced by $n-1$, $m-1$.

For $\pi\in S_{n,m}$, we obtain

$$(4.7)\qquad \Xi_n\left(\operatorname{sgn}(\pi)\prod_{i=1}^{n} t_i^{(n-\pi(i))+\chi(i\leqq m)}\right) = \det\left|t_j^{(n-i)+\chi(i\leqq m)}\right|_{n\times n}.$$

The result follows by using (4.7) and (3.8) to evaluate the contributions (4.5) to the left side of (4.4). □

Let $d_\mu(\mathbf{t})$ be an n-dimensional antisymmetric measure. That is, for all $\pi\in S_n$

$$(4.8)\qquad d_\mu(\pi(\mathbf{t})) = \operatorname{sgn}(\pi)\, d_\mu(\mathbf{t}).$$

Multiply both sides of (4.4) by $d_\mu(\mathbf{t})$ and integrate. The Jacobian of the transformation

$t_i \to t_{\pi(i)}$, $1 \leqq i \leqq n$, is sgn (π). Using (3.8), we obtain

$$\begin{aligned}(4.9)\qquad &\int \cdots \int \prod_{i=1}^{m} t_i \, {}_Q\Delta_n(\mathbf{t})\, d_\mu(\mathbf{t}) \\ &= \left[\sum_{\pi \in S_{n,m}} \prod_{\substack{1 \leqq i<j \leqq n \\ \pi(i)>\pi(j)}} Q_{i,j}\right] \int \cdots \int \prod_{i=1}^{n} t_i^{(n-i)+\chi(i \leqq m)}\, d_\mu(\mathbf{t}).\end{aligned}$$

When $Q_{i,j}=1$, $1 \leqq i<j \leqq n$, (4.9) becomes

$$(4.10)\qquad \int \cdots \int \prod_{i=1}^{m} t_i \Delta_n(\mathbf{t})\, d_\mu(\mathbf{t}) = m!(n-m)! \int \cdots \int \prod_{i=1}^{n} t_i^{(n-i)+\chi(i \leqq m)}\, d_\mu(\mathbf{t}).$$

Comparing (4.9) and (4.10), we obtain

$$\begin{aligned}(4.11)\qquad &\int \cdots \int \prod_{i=1}^{m} t_i \, {}_Q\Delta_n(\mathbf{t})\, d_\mu(\mathbf{t}) \\ &= \frac{1}{m!(n-m)!}\left[\sum_{\pi \in S_{n,m}} \prod_{\substack{1 \leqq i<j \leqq n \\ \pi(i)>\pi(j)}} Q_{i,j}\right] \int \cdots \int \prod_{i=1}^{m} t_i \Delta_n(\mathbf{t})\, d_\mu(\mathbf{t}).\end{aligned}$$

It is well known (see MacMahon [21]) that

$$(4.12)\qquad \sum_{\pi \in S_n} \prod_{\substack{1 \leqq i<j \leqq n \\ \pi(i)>\pi(j)}} Q = \prod_{i=1}^{n} \frac{(1-Q^i)}{(1-Q)}.$$

For $m=0$, (4.4) becomes

$$(4.13)\qquad \Xi_n\left(\prod_{1 \leqq i<j \leqq n} (t_i - Qt_j)\right) = \prod_{i=1}^{n} \frac{(1-Q^i)}{(1-Q)} \Delta_n(\mathbf{t}).$$

This is a result of Macdonald [17] for the root system A_{n-1}. See Carter [10, Thms. 10.2.1, 10.2.3] and Kadell [16].

Using the simple identity

$$(4.14)\qquad (\alpha)_n = (-\alpha)^n q^{\binom{n}{2}} \left(\frac{1}{\alpha q^{n-1}}\right)_n,$$

we find that

$$(4.15)\qquad t_i^{(2k-1)} \left(\frac{t_j q^{1-k}}{t_i}\right)_{2k-1} = -t_j^{(2k-1)} \left(\frac{t_i q^{1-k}}{t_j}\right)_{2k-1}.$$

Thus

$$(4.16)\qquad {}_q\Delta_n^{(2k-1)}(\mathbf{t}) = \prod_{1 \leqq i<j \leqq n} t_i^{(2k-1)} \left(\frac{t_j q^{1-k}}{t_i}\right)_{2k-1}$$

is an antisymmetric function and

$$(4.17)\qquad {}_q\Delta_n^{2k}(\mathbf{t}) = \Delta_n(\mathbf{t})\, {}_q\Delta_n^{(2k-1)}(\mathbf{t})$$

is a symmetric function.

Let $m=0$, $Q_{i,j}=q^k$, $1\leqq i<j\leqq n$, and choose $d_\mu(\mathbf{t})$ so that the left side of (4.11) is Askey's integral in (1.11). Using (4.11) and (4.12), we obtain

$$\begin{aligned}\int_0^1\cdots\int_0^1\prod_{i=1}^n t_i^{(x-1)}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\,{}_q\Delta_n^{2k}(\mathbf{t})\,d_qt_1\cdots d_qt_n\\ =n!\,q^{[kx\binom{n}{2}+2k^2\binom{n}{3}]}\prod_{i=1}^n\frac{\Gamma_q(x+(n-i)k)\Gamma_q(y+(n-i)k)\Gamma_q(ik)}{\Gamma_q(x+y+(2n-i-1)k)\Gamma_q(k)},\end{aligned}\tag{4.18}$$

which is the symmetric version of Askey's conjecture (1.11). Observe that the integrand in (4.18) is a symmetric function. This is the first step in Macdonald's proof [19] of (1.11) for $k=1$.

While we offer only the Laurent expansion (4.2) of ${}_Q\Delta_n(\mathbf{t})$, the ordinary Vandermonde $\Delta_n(\mathbf{t})$ is capable of many expansions. We have the following lemma.

LEMMA 5. *Let $B_0=\{p_1(t),p_2(t),\cdots,p_n(t)\}$ be a basis for the space of polynomials in t of degree at most $n-1$. Then there is a constant c such that*

$$\det|p_i(t_j)|_{n\times n}=c\Delta_n(\mathbf{t}).\tag{4.19}$$

Proof. We proceed by induction on s to show that

$$B_s=\{t^{n-1},\cdots,t^{n-s},p_{s+1}(t),\cdots,p_n(t)\}\tag{4.20}$$

is a basis for all s with $0\leqq s\leqq n$, provided the basis B_0 is properly ordered. This holds for $s=0$ by assumption. Let $0\leqq s\leqq n-1$ and write t^{n-s-1} in terms of the basis B_s. This gives

$$t^{n-s-1}=\sum_{r=1}^{s}c_{s,r}t^{n-r}+\sum_{r=s+1}^{n}c_{s,r}p_r(t).\tag{4.21}$$

Since $t^{n-s-1}\notin\operatorname{span}(t^{n-1},\cdots,t^{n-s})$, $c_{s,r}$ cannot be 0 for all r with $s+1\leqq r\leqq n$. We assume that B_0 is ordered so that

$$c_{s,s+1}\neq 0.\tag{4.22}$$

Solving (4.21) for $p_{s+1}(t)$, we see that $p_{s+1}(t)\in\operatorname{span}(B_{s+1})$ and hence B_{s+1} is a basis. For $0\leqq s\leqq n$, we set

$$D_s=\begin{vmatrix} t_1^{n-1} & t_2^{n-1} & \cdots & t_n^{n-1}\\ \vdots & \vdots & \cdots & \vdots\\ t_1^{n-s} & t_2^{n-s} & \cdots & t_n^{n-s}\\ p_{s+1}(t_1) & p_{s+1}(t_2) & \cdots & p_{s+1}(t_n)\\ \vdots & \vdots & \cdots & \vdots\\ p_n(t_1) & p_n(t_2) & \cdots & p_n(t_n)\end{vmatrix}.\tag{4.23}$$

Let $0\leqq s\leqq n-1$. Multiply row $s+1$ of D_s by $c_{s,s+1}$ and for each r with $1\leqq r\leqq n$, $r\neq s+1$, add $c_{s,r}$ times row r to row $s+1$. We obtain

$$\det(D_{s+1})=c_{s,s+1}\det(D_s),\qquad 0\leqq s\leqq n-1,\tag{4.24}$$

provided the basis B_0 has been properly ordered. The left side of (4.19) is $\det(D_0)$ and $\det(D_n)=\Delta_n(\mathbf{t})$. Thus (4.24) yields

$$\det|p_i(t_j)|_{n\times n}=\prod_{s=0}^{n-1}\frac{1}{c_{s,s+1}}\Delta_n(\mathbf{t}),\tag{4.25}$$

as required. □

Kadell [15, Lemma 7] gives a weaker version of Lemma 5. See Gordon and Houten [12a], [12b] for an interesting special case which is related to plane partitions. We require the special cases

$$\det|(At_j)_{(i-1)}|_{n\times n} = q^{\binom{n}{3}}A^{\binom{n}{2}}\Delta_n(\mathbf{t}), \tag{4.26}$$

$$\det|t_j^{n-i}(At_j)_{(i-1)}|_{n\times n} = \Delta_n(\mathbf{t}), \tag{4.27}$$

and

$$\det|(Aq^{n-i}t_j)_{(i-1)}|_{n\times n} = q^{2\binom{n}{3}}A^{\binom{n}{2}}\Delta_n(\mathbf{t}). \tag{4.28}$$

5. The case $n=2$, $l=0$. Let $L\subset\{1,\cdots,n\}$ and set

$$\begin{aligned} {}_Q\Delta_{n,L}(\mathbf{t}) &= \Delta_n(Q^{\chi(1\in L)}t_1,\cdots,Q^{\chi(n\in L)}t_n)\\ &= \prod_{1\leqq i<j\leqq n}(Q^{\chi(i\in L)}t_i - Q^{\chi(j\in L)}t_j)\\ &= Q^{[\sum_{i\in L}n-i]}\prod_{1\leqq i<j\leqq n}(t_i - Q^{\chi(j\in L)-\chi(i\in L)}t_j) \end{aligned} \tag{5.1}$$

and

$$\begin{aligned} {}_Q\Delta_{n,m,l}(\mathbf{t}) &= {}_Q\Delta_{n,\{1,\cdots,m\}\cup\{n-l+1,\cdots,n\}}(\mathbf{t})\\ &= \Delta_n(Qt_1,\cdots,Qt_m,t_{m+1},\cdots,t_{n-l},Qt_{n-l+1},\cdots,Qt_n). \end{aligned} \tag{5.2}$$

The support of the measure d_qt has an accumulation point at $t=0$, but not at $t=1$. We cannot make the substitution $t_i\to(1-t_{n-i+1})$, $1\leqq i\leqq n$, as we did for (1.3). This accounts for the asymmetry of the exponent of q in (1.11). Set

$$\begin{aligned} {}_qI_{n,m,l}(x,y,k) = \int_0^1\cdots\int_0^1\prod_{i=1}^n t_i^{(x-1)+\chi(i\leqq m)}\frac{(t_iq)_\infty}{(t_iq^{y+\chi(n-i+1\leqq l)})_\infty}\\ \cdot{}_{(q^k)}\Delta_{n,0,l}(\mathbf{t})\,{}_q\Delta_n^{(2k-1)}(\mathbf{t})\,d_qt_1\cdots d_qt_n. \end{aligned} \tag{5.3}$$

The q-multinomial coefficient is given by

$$\begin{bmatrix} A \\ a_1,\cdots,a_n \end{bmatrix}_q = \frac{(q)_A}{(q)_{a_1}\cdots(q)_{a_n}(q)_{(A-a_1-\cdots-a_n)}}. \tag{5.4}$$

In particular,

$$\begin{bmatrix} n \\ m \end{bmatrix}_q = \frac{(q^{n-m+1})_m}{(q)_m}, \qquad \begin{bmatrix} n \\ m,l \end{bmatrix}_q = \frac{(q^{n-m-l+1})_{m+l}}{(q)_m(q)_l}. \tag{5.5}$$

We have the following conjecture.

Conjecture 6.

$$\begin{aligned} {}_qI_{n,m,l}(x,y,k) = n!\frac{\begin{bmatrix} n \\ m,l \end{bmatrix}_{(q^k)}}{\binom{n}{m,l}}q^{[kx\binom{n}{2}+k\binom{m}{2}+k\binom{l}{2}+2k^2\binom{n}{3}]}\\ \cdot\prod_{i=1}^n\frac{\Gamma_q(x+(n-i)k+\chi(i\leqq m))\Gamma_q(y+(n-i)k+\chi(i\leqq l))\Gamma_q(ik)}{\Gamma_q(x+y+(2n-i-1)k+\chi(i\leqq m+l))\Gamma_q(k)}. \end{aligned} \tag{5.6}$$

When $m=l=0$ or $m=n$, $l=0$, or $m=0$, $l=n$, (5.6) reduces to (4.18) which is equivalent to Askey's conjecture (1.11).

Observe that the right side of (5.6) satisfies

$$q^{[ky\binom{n}{2}]}\,{}_qI_{n,m,l}(x,y,k)=q^{[kx\binom{n}{2}]}\,{}_qI_{n,l,m}(y,x,k). \tag{5.7}$$

We should be able to prove Conjecture 6 if we could establish the symmetry (5.7). We have the following theorem.

THEOREM 7. *Conjecture 6 holds when $n=2$ and $l=0$.*

Proof. Askey [6] has shown that (1.11) holds when $n=2$. Thus (4.18) holds when $n=2$ and (5.6) holds when $n=2$, $m=l=0$, and when $n=m=2$, $l=0$. We must treat the case $n=2$, $m=1$, $l=0$. Observe that

$$\frac{(t_iq)_\infty}{(t_iq^{y+1})_\infty}=\frac{(t_iq)_\infty}{(t_iq^y)_\infty}(1-t_iq^y) \tag{5.8}$$

and

$$(1-t_1q^y)(1-t_2q^y)=1-q^y(t_1+t_2)+q^{2y}t_1t_2. \tag{5.9}$$

Since ${}_{(q^k)}\Delta_{n,0,0}(\mathbf{t})\,{}_q\Delta_n^{(2k-1)}(\mathbf{t})={}_q\Delta_n^{2k}(\mathbf{t})$ is a symmetric function, we obtain

$${}_qI_{2,0}(x,y+1,k)={}_qI_{2,0}(x,y,k)-2q^y\,{}_qI_{2,1}(x,y,k)+q^{2y}\,{}_qI_{2,2}(x,y,k). \tag{5.10}$$

The result follows by solving (5.10) for ${}_qI_{2,1}(x,y,k)$. □

The case $n=2$, $m=1$, $l=0$, provides an interesting summation formula for a nearly well-poised ${}_3\phi_2$. We have the well-known q-binomial theorem (see Andrews [2, (2.2.1)])

$${}_1\phi_0\left[\begin{matrix}a\\ \end{matrix}\middle|\, q,t\right]=\frac{(at)_\infty}{(t)_\infty},\qquad |t|<1, \tag{5.11}$$

where the basic hypergeometric series is given by

$${}_{s+1}\phi_s\left[\begin{matrix}a_1,a_2,\cdots,a_{s+1}\\ b_1,b_2,\cdots,b_s\end{matrix}\middle|\, q,x\right]=\sum_{n=0}^{\infty}\frac{(a_1)_n\cdots(a_{s+1})_n}{(b_1)_n\cdots(b_s)_n}\frac{x^n}{(q)_n}. \tag{5.12}$$

Using the special case

$$t_1^{2k}\left(\frac{t_2q^{1-k}}{t_1}\right)_{2k}=\sum_{i=0}^{2k}\frac{(q^{-2k})_i}{(q)_i}q^{(1+k)i}t_1^{(2k-i)}t_2^i \tag{5.13}$$

of (5.11) and (1.10), we obtain

$$\begin{aligned}
{}_qI_{2,1}(x,y,k)&=\int_0^1\cdots\int_0^1 t_1^xt_2^{x-1}\frac{(t_1q)_\infty}{(t_1q^y)_\infty}\frac{(t_2q)_\infty}{(t_2q^y)_\infty}t_1^{2k}\left(\frac{t_2q^{1-k}}{t_1}\right)_{2k}d_qt_1\,d_qt_2\\
&=\sum_{i=0}^{2k}\frac{(q^{-2k})_i}{(q)_i}q^{(1+k)i}\int_0^1 t_1^{x+2k-i}\frac{(t_1q)_\infty}{(t_1q^y)_\infty}d_qt_1\\
&\quad\cdot\int_0^1 t_2^{x+2k-i}\frac{(t_2q)_\infty}{(t_2q^y)_\infty}d_qt_2\\
&=\sum_{i=0}^{2k}\frac{(q^{-2k})_i}{(q)_i}q^{(1+k)i}\frac{\Gamma_q(x+2k-i+1)\Gamma_q(y)}{\Gamma_q(x+y+2k-i+1)}\frac{\Gamma_q(x+i)\Gamma_q(y)}{\Gamma_q(x+y+i)}\\
&=\frac{\Gamma_q(x)\Gamma_q(x+2k+1)\Gamma_q(y)\Gamma_q(y)}{\Gamma_q(x+y)\Gamma_q(x+y+2k+1)}\,{}_3\phi_2\left[\begin{matrix}q^{-2k},q^{-2k-x-y},q^x\\ q^{-2k-x},q^{x+y}\end{matrix}\middle|\, q^{y+k+1}\right].
\end{aligned} \tag{5.14}$$

The first equality in (5.14) holds since we have added an antisymmetric function to the integrand. By Theorem 7, we have

$$(5.15)\qquad {}_qI_{2,1}(x,y,k)=q^{kx}(1+q^k)\frac{\Gamma_q(x)\Gamma_q(x+k+1)\Gamma_q(y)\Gamma_q(y+k)\Gamma_q(2k)}{\Gamma_q(x+y+k)\Gamma_q(x+y+2k+1)\Gamma_q(k)}.$$

Equating (5.14) and (5.15) and solving for the ${}_3\phi_2$ yields

$$(5.16)\qquad \begin{aligned}&{}_3\phi_2\left[\begin{matrix}q^{-2k},\,q^{-2k-x-y},\,q^{x}\\ q^{-2k-x},\,q^{x+y}\end{matrix}\,\middle|\,q^{y+k+1}\right]\\ &\quad=q^{kx}(1+q^k)\frac{\Gamma_q(x+k+1)\Gamma_q(y+k)\Gamma_q(x+y)\Gamma_q(2k)}{\Gamma_q(x+2k+1)\Gamma_q(y)\Gamma_q(x+y+k)\Gamma_q(k)}.\end{aligned}$$

This may be written as

$$(5.17)\qquad {}_3\phi_2\left[\begin{matrix}q^{-2k},\,a,\,b\\ q^{-2k}/a,\,q^{-2k}/b\end{matrix}\,\middle|\,q,\,q^{1-k}/ab\right]=\frac{(q^{k+1})_k(abq^{k+1})_k}{(aq^{k+1})_k(bq^{k+1})_k}.$$

Compare this with

$$(5.18)\qquad {}_3\phi_2\left[\begin{matrix}q^{-2k},\,a,\,b\\ q^{1-2k}/a,\,q^{1-2k}/b\end{matrix}\,\middle|\,q,\,q^{2-k}/ab\right]=\frac{(q^{k+1})_k(abq^{k})_k}{(aq^{k})_k(bq^{k})_k},$$

which is due to Jackson [14]. Carlitz [9] gives another proof of (5.18) using some q-analogues of quadratic transformations.

6. A proof of Conjecture 6 for $k=1$, $l=0$. Although our q-integral is a discrete sum, the integrand vanishes if any two of the variables $t_1, t_2, \cdots, t_n$ are equal. As with (2.1), we obtain

$$(6.1a)\qquad \begin{aligned}{}_qI_{n,m}(x,y,k)=m\int_0^1 t_1^x\frac{(t_1q)_\infty}{(t_1q^y)_\infty}\Bigg[\int_{t_1}^1\cdots\int_{t_1}^1\prod_{i=2}^{n}t_i^{(x-1)+\chi(i\leqq m)}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\\ \cdot{}_q\Delta_n^{2k}(\mathbf{t})\,d_qt_2\cdots d_qt_n\Bigg]\,d_qt_1,\end{aligned}$$

$$(6.1b)\qquad \begin{aligned}+(n-m)\int_0^1 t_n^{(x-1)}\frac{(t_nq)_\infty}{(t_nq^y)_\infty}\Bigg[\int_{t_n}^1\cdots\int_{t_n}^1\prod_{i=1}^{n-1}t_i^{(x-1)+\chi(i\leqq m)}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\\ \cdot{}_q\Delta_n^{2k}(\mathbf{t})\,d_qt_1\cdots d_qt_{n-1}\Bigg]\,d_qt_n.\end{aligned}$$

It is easy to use the q-beta integral (1.10) to obtain a recurrence relation as we did in § 2. Equation (1.9) gives

$$(6.2)\qquad \lim_{x\to0}\frac{(1-q^x)}{(1-q)}\Gamma_q(x)=\lim_{x\to0}\Gamma_q(x+1)=1.$$

Under the hypotheses of Lemma 3, we obtain

$$(6.3)\qquad \lim_{x\to0}\frac{(1-q^x)}{(1-q)}\int_0^1 t^{x-1}\frac{(tq)_\infty}{(tq^y)_\infty}f(x,t)\,d_qt=f(0,0).$$

In replacing (2.2), we must change d_qt to $d_{(q^x)}u$. This shows that (6.1a) contributes 0 to the limit in (6.4) below. Taking $f(x,t)$ to be an $(n-1)$-dimensional q-integral as in (2.6), we obtain

$$(6.4)\qquad \lim_{x\to0}\frac{(1-q^x)}{(1-q)}{}_qI_{n,m}(x,y,k)=(n-m)\,{}_qI_{n-1,m}(2k,y,k).$$

This is valid for all m, $0 \leqq m \leqq n$, using the convention

$$ {}_qI_{n-1,n}(x, y, k) = {}_qI_{n-1,n-1}(x, y, k). \tag{6.5} $$

The cancellation argument for (3.6) gives

$$ \begin{aligned} {}_qI_{n,m}(x, y, k) = m!(n-m)! \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i\leqq m)+(n-i)} \frac{(t_iq)_\infty}{(t_iq^y)_\infty} \\ \cdot {}_q\Delta_n^{(2k-1)}(\mathbf{t})\, d_qt_1 \cdots d_qt_n. \end{aligned} \tag{6.6} $$

Setting $A = q^y$ in (4.26) and solving for $\Delta_n(\mathbf{t})$, we obtain

$$ \Delta_n(\mathbf{t}) = q^{-[y\binom{n}{2}+\binom{n}{3}]} \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} (t_iq^y)_{(\pi(i)-1)}. \tag{6.7} $$

For $k = 1$, we expand the Vandermonde in (6.6) using (6.7). This yields

$$ \begin{aligned} {}_qI_{n,m}(x, y, 1) = m!(n-m)!\, q^{-[y\binom{n}{2}+\binom{n}{3}]} \\ \cdot \prod_{i=1}^{n} \Gamma_q(x+n-i+\chi(i\leqq m))\Gamma_q(y+n-i)\ {}_qU_{n,m}(x, y, 1), \end{aligned} \tag{6.8} $$

where

$$ {}_qU_{n,m}(x, y, 1) = \sum_{\pi\in S_n} \operatorname{sgn}(\pi) \prod_{i=1}^{n} \frac{1}{\Gamma_q(x+y+n-i+\pi(i)-1+\chi(i\leqq m))}. \tag{6.9} $$

Since ${}_qU_{n,m}(x, y, 1)$ is a function of $x+y$ rather than of x and y, we have

$$ {}_qU_{n,m}(x, y, 1) = {}_qU_{n,m}(0, x+y, 1). \tag{6.10} $$

We evaluate ${}_qU_{n,m}(x, y, 1)$ using the argument of § 3. We proceed by induction on n and let $0 \leqq m < n$. By (6.8) we obtain

$$ \begin{aligned} \lim_{x\to 0} \frac{(1-q^x)}{(1-q)}\, {}_qI_{n,m}(x, y, 1) = m!(n-m)!\, q^{-[y\binom{n}{2}+\binom{n}{3}]} \\ \cdot \Gamma_q(y) \prod_{i=1}^{n-1} \Gamma_q(n-i+\chi(i\leqq m))\Gamma_q(y+n-i)\ {}_qU_{n,m}(0, y, 1). \end{aligned} \tag{6.11} $$

The recurrence relation (6.4) and our induction hypothesis give

$$ \begin{aligned} &\lim_{x\to 0} \frac{(1-q^x)}{(1-q)}\, {}_qI_{n,m}(x, y, 1) \\ &\quad = (n-m)\ {}_qI_{n-1,m}(2, y, 1) \\ &\quad = (n-m)(n-1)! \frac{\begin{bmatrix} n-1 \\ m \end{bmatrix}_q}{\binom{n-1}{m}} q^{[\binom{m}{2}+2\binom{n}{3}]} \\ &\qquad \cdot \prod_{i=1}^{n-1} \frac{\Gamma_q(n-i+1+\chi(i\leqq m))\Gamma_q(y+n-i-1)}{\Gamma_q(y+2n-i-1+\chi(i\leqq m))} \Gamma_q(i). \end{aligned} \tag{6.12} $$

Solving (6.11) and (6.12) for ${}_qU_{n,m}(0, y, 1)$, we obtain

$$
\begin{aligned}
{}_qU_{n,m}(0,y,1) &= \begin{bmatrix} n-1 \\ m \end{bmatrix}_q q^{[y\binom{n}{2}+\binom{m}{2}+3\binom{n}{3}]} \frac{1}{\Gamma_q(y)} \\
&\quad \cdot \prod_{i=1}^{n-1} \frac{(1-q^{n-i+\chi(i\leqq m)})}{(1-q^{y+n-i-1})} \frac{\Gamma_q(i)}{\Gamma_q(y+2n-i-1+\chi(i\leqq m))} \\
&= \begin{bmatrix} n \\ m \end{bmatrix}_q q^{[y\binom{n}{2}+\binom{m}{2}+3\binom{n}{3}]} \frac{1}{\Gamma_q(y)} \frac{(1-q^{n-m})}{(1-q^n)} \\
&\quad \cdot \prod_{i=1}^{n-1} \frac{(1-q^{n-i+\chi(i\leqq m)})}{(1-q^{y+n-i-1})} \frac{\Gamma_q(i)}{\Gamma_q(y+2n-i-1+\chi(i\leqq m))} \\
&= \begin{bmatrix} n \\ m \end{bmatrix}_q q^{[y\binom{n}{2}+\binom{m}{2}+3\binom{n}{3}]} \prod_{i=1}^{n} \frac{\Gamma_q(i)}{\Gamma_q(y+2n-i-1+\chi(i\leqq m))}.
\end{aligned} \tag{6.13}
$$

Replacing y by $x+y$ in (6.13) and using (6.10) yields

$$
{}_qU_{n,m}(x,y,1) = \begin{bmatrix} n \\ m \end{bmatrix}_q q^{[x\binom{n}{2}+y\binom{n}{2}+\binom{m}{2}+3\binom{n}{3}]} \prod_{i=1}^{n} \frac{\Gamma_q(i)}{\Gamma_q(x+y+2n-i-1+\chi(i\leqq m))}. \tag{6.14}
$$

Substituting (6.14) into (6.8) gives

$$
\begin{aligned}
{}_qI_{n,m}(x,y,1) &= n! \frac{\begin{bmatrix} n \\ m \end{bmatrix}_q}{\binom{n}{m}} q^{[x\binom{n}{2}+\binom{m}{2}+2\binom{n}{3}]} \\
&\quad \cdot \prod_{i=1}^{n} \frac{\Gamma_q(x+n-i+\chi(i\leqq m))\Gamma_q(y+n-i)}{\Gamma_q(x+y+2n-i-1+\chi(i\leqq m))} \Gamma_q(i),
\end{aligned} \tag{6.15}
$$

in agreement with Conjecture 6 for $k=1$, $l=0$.

7. A proof of a q-analogue of Conjecture 2 for $k=1$. A q-analogue of Conjecture 2 is given by the following conjecture.

Conjecture 8. Let $\boldsymbol{\lambda}$ denote a partition $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_n \geqq 0$ with at most n parts. For all $k \geqq 0$ there exists a homogeneous symmetric polynomial ${}_qs^k_{n,\boldsymbol{\lambda}}(\mathbf{t})$ with leading term $\prod_{i=1}^n t_i^{\lambda_i}$ such that

$$
\begin{aligned}
{}_qI_{n,\boldsymbol{\lambda}}(x,y,k) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)} \frac{(t_iq)_\infty}{(t_iq^y)_\infty} {}_qs^k_{n,\boldsymbol{\lambda}}(\mathbf{t})\, {}_q\Delta_n^{2k}(\mathbf{t})\, d_qt_1 \cdots d_qt_n \\
&= n!\, q^{[k(\sum_{i=1}^n (i-1)\lambda_i)+kx\binom{n}{2}+2k^2\binom{n}{3}]} {}_qf^k_{n,\boldsymbol{\lambda}} \\
&\quad \cdot \prod_{i=1}^{n} \frac{\Gamma_q(x+(n-i)k+\lambda_i)\Gamma_q(y+(n-i)k)}{\Gamma_q(x+y+(2n-i-1)k+\lambda_i)},
\end{aligned} \tag{7.1}
$$

where

$$
{}_qf^k_{n,\boldsymbol{\lambda}} = \prod_{1\leqq i<j\leqq n} \frac{(q^{k(j-i)+\lambda_i-\lambda_j})_k}{(1-q)^k}. \tag{7.2}
$$

We now show that Conjecture 8 holds for $k=1$ with

$$
{}_qs^1_{n,\boldsymbol{\lambda}}(\mathbf{t}) = s^1_{n,\boldsymbol{\lambda}}(\mathbf{t}). \tag{7.3}
$$

Since the integrand is a symmetric function and contains $\Delta_n(\mathbf{t})$ as a factor, the argument for (2.15) and (6.3) yields

$$\lim_{x\to 0}\frac{(1-q^x)}{(1-q)}\,{}_qI_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(x,y,1)=n\,{}_qI_{n-1,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1})}(2,y,1). \tag{7.4}$$

Substituting (3.21) into (7.1) gives

$${}_qI_{n,\boldsymbol{\lambda}}(x,y,1)=n!\int_0^1\cdots\int_0^1\prod_{i=1}^n t_i^{(x-1)+(n-i+\lambda_i)}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\Delta_n(\mathbf{t})\,d_qt_1\cdots d_qt_n. \tag{7.5}$$

Using (6.7) to expand the Vandermonde in (7.5) yields

$${}_qI_{n,\boldsymbol{\lambda}}(x,y,1)=n!\,q^{-[y\binom{n}{2}+\binom{n}{3}]}\prod_{i=1}^n\Gamma_q(x+n-i+\lambda_i)\Gamma_q(y+n-i)\,{}_qU_{n,\boldsymbol{\lambda}}(x,y,1), \tag{7.6}$$

where

$${}_qU_{n,\boldsymbol{\lambda}}(x,y,1)=\sum_{\pi\in S_n}\operatorname{sgn}(\pi)\prod_{i=1}^n\frac{1}{\Gamma_q(x+y+n-i+\lambda_i+\pi(i)-1)}. \tag{7.7}$$

We have

$${}_qU_{n,\boldsymbol{\lambda}}(x,y,1)={}_qU_{n,(\lambda_1-\lambda_n,\cdots,\lambda_{n-1}-\lambda_n,0)}(0,x+y+\lambda_n,1). \tag{7.8}$$

We proceed by induction on n. We treat first the case $\lambda_n=0$. By (7.6) we have

$$\begin{aligned}&\lim_{x\to 0}\frac{(1-q^x)}{(1-q)}\,{}_qI_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(x,y,1)\\&\quad=n!\,q^{-[y\binom{n}{2}+\binom{n}{3}]}\Gamma_q(y)\prod_{i=1}^{n-1}\Gamma_q(n-i+\lambda_i)\Gamma_q(y+n-i)\\&\qquad\cdot{}_qU_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(0,y,1).\end{aligned} \tag{7.9}$$

Our recurrence relation (7.4) and our induction hypothesis give

$$\begin{aligned}&\lim_{x\to 0}\frac{(1-q^x)}{(1-q)}\,{}_qI_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(x,y,1)\\&\quad=n(n-1)!\,q^{[\sum_{i=1}^{n-1}(i-1)\lambda_i+2\binom{n-1}{2}+2\binom{n-1}{3}]}\,{}_qf^1_{n-1,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1})}\\&\qquad\cdot\prod_{i=1}^{n-1}\frac{\Gamma_q(n-i+1+\lambda_i)\Gamma_q(y+n-i-1)}{\Gamma_q(y+2n-i-1+\lambda_i)}\\&\quad=n!\,q^{[\sum_{i=1}^{n-1}(i-1)\lambda_i+2\binom{n}{3}]}\prod_{1\le i<j\le n-1}\frac{(1-q^{j-i+\lambda_i-\lambda_j})}{(1-q)}\\&\qquad\cdot\prod_{i=1}^{n-1}\frac{\Gamma_q(n-i+1+\lambda_i)\Gamma_q(y+n-i-1)}{\Gamma_q(y+2n-i-1+\lambda_i)}.\end{aligned} \tag{7.10}$$

Solving (7.9) and (7.10) for ${}_qU_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(0,y,1)$ yields

$$\begin{aligned}&{}_qU_{n,(\lambda_1,\lambda_2,\cdots,\lambda_{n-1},0)}(0,y,1)\\&\quad=q^{[\sum_{i=1}^{n-1}(i-1)\lambda_i+y\binom{n}{2}+3\binom{n}{3}]}\prod_{1\le i<j\le n-1}\frac{(1-q^{j-i+\lambda_i-\lambda_j})}{(1-q)}\\&\qquad\cdot\prod_{i=1}^{n-1}\frac{(1-q^{n-i+\lambda_i})}{(1-q)}\prod_{i=1}^n\frac{1}{\Gamma_q(y+2n-i-1+\lambda_i)}.\end{aligned} \tag{7.11}$$

Replacing y by $x+y+\lambda_n$ and λ_i, $1\leqq i\leqq n-1$, by $\lambda_i-\lambda_n$, $1\leqq i\leqq n-1$, in (7.11) and using (7.8) yields

$$
\begin{aligned}
{}_qU_{n,\boldsymbol{\lambda}}(x,y,1) &= q^{[\sum_{i=1}^n (i-1)\lambda_i + x\binom{n}{2} + y\binom{n}{2} + 3\binom{n}{3}]} \\
&\quad \cdot \prod_{1\leqq i<j\leqq n} \frac{(1-q^{j-i+\lambda_i-\lambda_j})}{(1-q)} \prod_{i=1}^n \frac{1}{\Gamma_q(x+y+2n-i-1+\lambda_i)}.
\end{aligned} \tag{7.12}
$$

Substituting (7.12) into (7.6), we get the required result (7.1).

8. The case $l>0$. The following special case of Lemma 4 is important enough to be labeled a theorem.

THEOREM 9.

$$
\sum_{\substack{M\subset\{1,\cdots,n\},|M|=m \\ L\subset\{1,\cdots,n\}-M,|L|=l \\ J\subset L,|J|=j}} \prod_{i\in M\cup J} t_i\, {}_Q\Delta_{n,L}(\mathbf{t}) = {}_Qc_{n,m,l,j}\, e_{n,m+j}(\mathbf{t})\Delta_n(\mathbf{t}), \tag{8.1}
$$

where

$$
{}_Qc_{n,m,l,j} = Q^{[(n-ml)j+\binom{l}{2}]} \begin{bmatrix} m+j \\ j \end{bmatrix}_Q \begin{bmatrix} n-m-j \\ l-j \end{bmatrix}_Q. \tag{8.2}
$$

Proof. It is easy to see that since $\Delta_n(\mathbf{t})$ is an antisymmetric function, so is the left side of (8.1). Explicitly, $m!(n-m-l)!j!(l-j)!$ times the left side of (8.1) equals

$$
\Xi_n\left(\prod_{i=1}^{m+j} t_i\, {}_Q\Delta_{n,\{m+1,\cdots,m+l\}}(\mathbf{t})\right). \tag{8.3}
$$

Thus (8.1) holds for some constant ${}_Qc_{n,m,l,j}$, which can be evaluated using (5.1) and (4.4). Alternatively, we set

$$
t_i = Q^i, \qquad 1\leqq i\leqq n. \tag{8.4}
$$

The left side of (8.1) vanishes unless $L=\{n-l+1,\cdots,n\}$. Thus

$$
\begin{aligned}
&\left(\sum_{\substack{M\subset\{1,\cdots,n-l\} \\ |M|=m}} Q^{[\sum_{i\in M} i]}\right)\left(\sum_{\substack{J\subset\{n-l+1,\cdots,n\} \\ |J|=j}} Q^{[\sum_{i\in J} i]}\right)\Delta_n(Q,\cdots,Q^{n-l},Q^{n-l+2},\cdots,Q^{n+1}) \\
&\qquad = {}_Qc_{n,m,l,j}\left(\sum_{\substack{S\subset\{1,\cdots,n\} \\ |S|=m+j}} Q^{[\sum_{i\in S} i]}\right)\Delta_n(Q,\cdots,Q^n).
\end{aligned} \tag{8.5}
$$

The case $a=q^{-N}$, $t=q^{N+s}x$, of the q-binomial theorem (5.11) is

$$
(q^s x)_N = \sum_{n=0}^N q^{[sn+\binom{n}{2}]} \begin{bmatrix} N \\ n \end{bmatrix}_q (-x)^n. \tag{8.6}
$$

Replacing q by Q and equating coefficients of x^n yields

$$
\sum_{\substack{A\subset\{s,\cdots,s+N-1\} \\ |A|=n}} Q^{[\sum_{i\in S} i]} = Q^{[sn+\binom{n}{2}]} \begin{bmatrix} N \\ n \end{bmatrix}_Q. \tag{8.7}
$$

Using (8.7) to evaluate the three sums in (8.5) and solving for ${}_Qc_{n,m,l,j}$, we obtain (8.2), as required. □

The parameter $l\geqq 0$ introduces the factor

$$
\prod_{i=n-l+1}^n (1-At_i) = \sum_{j=0}^l (-A)^j e_{l,j}(t_{n-l+1},\cdots,t_n) \tag{8.8}
$$

with $A = q^y$ into the integrand in (5.3) and replaces the Vandermonde $\Delta_n(\mathbf{t})$ by ${}_{(q^k)}\Delta_{n,0,l}(\mathbf{t})$ (recall (5.2)). Let $d_\mu(\mathbf{t})$ be an n-dimensional antisymmetric measure. We have

$$\begin{aligned}&\int\cdots\int \prod_{i=1}^{m} t_i \prod_{i=n-l+1}^{n} (1-At_i)\ {}_Q\Delta_{n,0,l}(\mathbf{t})\ d_\mu(\mathbf{t})\\ &\quad=\sum_{j=0}^{l}(-A)^j\binom{l}{j}\int\cdots\int \prod_{i=1}^{m} t_i \prod_{i=n-j+1}^{n} t_i\ {}_Q\Delta_{n,0,l}(\mathbf{t})\ d_\mu(\mathbf{t}),\end{aligned}\tag{8.9}$$

since each of the $\binom{l}{j}$ terms in the expansion (3.7) of the elementary symmetric function in (8.8) gives the same contribution to the integral. Multiply both sides of (8.1) by $d_\mu(\mathbf{t})$ and integrate. The $\binom{n}{m,l}\binom{l}{j}$ terms on the left side of (8.1) give the same contribution to the integral as do the $\binom{n}{m+j}$ terms that arise when the elementary symmetric function on the right side is expanded. We obtain

$$\begin{aligned}&\binom{n}{m,l}\binom{l}{j}\int\cdots\int \prod_{i=1}^{m} t_i\ {}_Q\Delta_{n,0,l}(\mathbf{t})\ d_\mu(\mathbf{t})\\ &\quad={}_Qc_{n,m,l,j}\binom{n}{m+j}\int\cdots\int \prod_{i=1}^{m+j} t_i\ \Delta_n(\mathbf{t})\ d_\mu(\mathbf{t}).\end{aligned}\tag{8.10}$$

Solving (8.10) for the integral on the left side and substituting into (8.9) yields

$$\begin{aligned}&\int\cdots\int \prod_{i=1}^{m} t_i \prod_{i=n-l+1}^{n} (1-At_i)\ {}_Q\Delta_{n,0,l}(\mathbf{t})\ d_\mu(\mathbf{t})\\ &\quad=\frac{1}{\binom{n}{m,l}}\sum_{j=0}^{l}(-A)^j\,{}_Qc_{n,m,l,j}\binom{n}{m+j}\int\cdots\int \prod_{i=1}^{m+j} t_i\,\Delta_n(\mathbf{t})\ d_\mu(\mathbf{t}).\end{aligned}\tag{8.11}$$

Setting $Q = q^k$, $A = q^y$, and taking the appropriate choice of $d_\mu(\mathbf{t})$, (8.11) becomes

$${}_qI_{n,m,l}(x,y,k)=\frac{1}{\binom{n}{m,l}}\sum_{j=0}^{l}(-q^y)^j\,{}_{(q^k)}c_{n,m,l,j}\binom{n}{m+j}\,{}_qI_{n,m+j}(x,y,k).\tag{8.12}$$

We have the following theorem.

THEOREM 10. *Fix $n \geqq 2$, x, y and $k \geqq 0$. Conjecture* 6 (5.6) *holds for all $m \geqq 0$, $l \geqq 0$, $m+l \leqq n$, if and only if it holds for all m, $0 \leqq m \leqq n$, when $l = 0$.*

Proof. We must show that (8.12) holds when we substitute (5.6). Dividing by

$$\frac{1}{\binom{n}{m,l}}\,{}_{(q^k)}c_{n,m,l,0}\binom{n}{m}\,{}_qI_{n,m}(x,y,k),$$

this is

$$\frac{(q^{y+(n-l)k};q^k)_l}{(q^{x+y+(2n-m-l-1)k};q^k)_l}=\sum_{j=0}^{l}\frac{(q^{-kl};q^k)_j(q^{-x-(n-m-1)k};q^k)_j}{(q^k;q^k)_j(q^{-x-y-(2n-m-2)k};q^k)_j}\,q^{kj}.\tag{8.13}$$

This is equivalent to the special case (Slater [28, (3.3.2.7)])

$$\frac{(d/b)_n}{(d)_n}\,b^n=\sum_{i=0}^{n}\frac{(q^{-n})_i(b)_i}{(q)_i(d)_i}\,q^i\tag{8.14}$$

of the well-known q-analogue (Slater [28, (3.3.2.2)]) of Saalschütz's theorem. □

We have thus proved Conjecture 6 when $n=2$ and when $k=1$.

An alternative special case of the q-analogue of Saalschütz's theorem is the q-analogue (Andrews [2, Cor. 2.4]) of Gauss' theorem. We require only the terminating case (Slater [28, (3.3.2.6)])

$$\frac{(d/b)_n}{(d)_n}=\sum_{i=0}^{n}\frac{(q^{-n})_i(b)_i}{(q)_i(d)_i}\left(q^n\frac{d}{b}\right)^i. \tag{8.15}$$

Set

$$\begin{aligned}{}_qJ_{n,m,l}(x,y,k)=&\int_0^1\cdots\int_0^1\prod_{i=1}^{n}t_i^{(x-1)+\chi(i\leqq m)}\frac{(t_iq^{\chi(i\leqq n-l)})_\infty}{(t_iq^y)_\infty}\\&\cdot{}_{(q^{-k})}\Delta_{n,m,l}(\mathbf{t})\,{}_q\Delta_n^{(2k-1)}(\mathbf{t})\,d_qt_1\cdots d_qt_n.\end{aligned} \tag{8.16}$$

We have the following conjecture.

Conjecture 11.

$$\begin{aligned}{}_qJ_{n,m,l}(x,y,k)=&\,n!\frac{\begin{bmatrix}n\\m,\,l\end{bmatrix}_{(q^k)}}{\binom{n}{m,\,l}}q^{[kx\binom{n}{2}+xl+k\binom{m}{2}+k\binom{l}{2}+2k^2\binom{n}{3}]}\\&\cdot\prod_{i=1}^{n}\frac{\Gamma_q(x+(n-i)k+\chi(i\leqq m))\Gamma_q(y+(n-i)k+\chi(i\leqq l))}{\Gamma_q(x+y+(2n-i-1)k+\chi(i\leqq m+l))}\frac{\Gamma_q(ik)}{\Gamma_q(k)}.\end{aligned} \tag{8.17}$$

We may take $Q=q^{-k}$, $A=1$, in (8.11). Substituting into (8.17), we obtain an identity which is equivalent to (8.15) with base q^k. This proves the following theorem.

THEOREM 12. *Fix $n\geqq 2$, x, y and $k\geqq 0$. Conjecture* 11 (8.17) *holds for all $m\geqq 0$, $l\geqq 0$, $m+l\leqq n$ if and only if it holds for all m, $0\leqq m\leqq n$, when $l=0$.*

Since our formulas (5.6) and (8.17) for ${}_qI_n$ and ${}_qJ_n$ agree when $l=0$, they are equivalent by Theorems 10 and 12.

9. Further extensions of Selberg's integral. Following Askey [6], we may give q-analogues of Selberg's integral using a number of known q-analogues of the beta integral. Let

$$\int_0^\infty f(t)\,d_qt=(1-q)\sum_{n=-\infty}^{\infty}q^nf(q^n). \tag{9.1}$$

We may use

$$\int_0^\infty t^{(x-1)}\frac{(-cq^{x+y}t)_\infty}{(-ct)_\infty}\,d_qt=\frac{(-cq^x)_\infty}{(-c)_\infty}\frac{(-c^{-1}q^{1-x})_\infty}{(-c^{-1}q)_\infty}\frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}, \tag{9.2}$$

where there are no zero factors in the denominator of the integrand. This is given by Askey [5]. Set

$$\begin{aligned}{}_q^cF_{n,m,l}(x,y,k)=&\int_0^\infty\cdots\int_0^\infty\prod_{i=1}^{n}t_i^{(x-1)+\chi(i\leqq m)}\frac{(-cq^{x+y+2(n-1)k+\chi(i\leqq m)+\chi(n-i+1\leqq l)}t_i)_\infty}{(-ct_i)_\infty}\\&\cdot{}_{(q^k)}\Delta_{n,m,l}(\mathbf{t})\,{}_q\Delta_n^{(2k-1)}(\mathbf{t})\,d_qt_1\cdots d_qt_n\end{aligned} \tag{9.3}$$

and

$$\begin{aligned}{}_q^cG_{n,m,l}(x,y,k)=&\int_0^\infty\cdots\int_0^\infty\prod_{i=1}^{n}t_i^{(x-1)+\chi(i\leqq m)}\frac{(-cq^{x+y+2(n-1)k}t_i)_\infty}{(-cq^{-1+\chi(m<i\leqq n-l)}t_i)_\infty}\\&\cdot{}_{(q^{-k})}\Delta_{n,m,l}(\mathbf{t})\,{}_q\Delta_n^{(2k-1)}(\mathbf{t})\,d_qt_1\cdots d_qt_n.\end{aligned} \tag{9.4}$$

Since $k \geqq 0$ is a nonnegative integer, all of our q-analogues of $\Delta_n^{2k}(\mathbf{t})$ are polynomials. Thus, ${}_qF_n$ and ${}_qG_n$ divided by the nth power of (9.2) are rational functions in q, q^x, q^y and c. Similar remarks apply to ${}_qI_n$ and ${}_qJ_n$. For any integer μ, we have

$$
\begin{aligned}
&{}_{-q^{(\mu-x-y-2(n-1)k)}}F_{n,m,l}(x, y, k) \\
&\quad = q^{-[\mu(nx+m)+(2\mu-1)k\binom{n}{2}]}\, {}_qJ_{n,m,n-m-l}(x, -x-y-2(n-1)k, k)
\end{aligned} \tag{9.5}
$$

and

$$
\begin{aligned}
&{}_{-q^{(\mu-x-y-2(n-1)k)}}G_{n,m,l}(x, y, k) \\
&\quad = q^{-[(\mu-1)(nx+m)+(2\mu-1)k\binom{n}{2}]}\, {}_qI_{n,m,n-m-l}(x, -x-y-2(n-1)k, k).
\end{aligned} \tag{9.6}
$$

Since a rational function is determined by finitely many values, we see that Conjecture 6 (5.6) and Conjecture 11 (8.17) are equivalent to

$$
\begin{aligned}
{}_q^cF_{n,m,l}(x, y, k) &= n! \frac{\begin{bmatrix} n \\ m, l \end{bmatrix}_{(q^k)}}{\binom{n}{m, l}} q^{[km(n-1)-k\binom{m}{2}-k\binom{l}{2}]} \\
&\quad \cdot \prod_{i=1}^{n} \frac{(-cq^{x+2(n-i)k+\chi(i\leqq m)})_\infty}{(-c)_\infty} \frac{(-c^{-1}q^{1-x-2(n-i)k-\chi(i\leqq m)})_\infty}{(-c^{-1}q)_\infty} \\
&\quad \cdot \prod_{i=1}^{n} \frac{\Gamma_q(x+(n-i)k+\chi(i\leqq m))\Gamma_q(y+(n-i)k+\chi(i\leqq l))}{\Gamma_q(x+y+(2n-i-1)k+\chi(i\leqq m+l))} \frac{\Gamma_q(ik)}{\Gamma_q(k)}
\end{aligned} \tag{9.7}
$$

and

$$
\begin{aligned}
{}_q^cG_{n,m,l}(x, y, k) &= n! \frac{\begin{bmatrix} n \\ m, l \end{bmatrix}_{(q^k)}}{\binom{n}{m, l}} q^{[m(x+1)+xl+km(n-1)-k\binom{m}{2}-k\binom{l}{2}]} \\
&\quad \cdot \prod_{i=1}^{n} \frac{(-cq^{x+2(n-i)k+\chi(i\leqq m)})_\infty}{(-c)_\infty} \frac{(-c^{-1}q^{1-x-2(n-i)k-\chi(i\leqq m)})_\infty}{(-c^{-1}q)_\infty} \\
&\quad \cdot \prod_{i=1}^{n} \frac{\Gamma_q(x+(n-i)k+\chi(i\leqq m))\Gamma_q(y+(n-i)k+\chi(i\leqq l))}{\Gamma_q(x+y+(2n-i-1)k+\chi(i\leqq m+l))} \frac{\Gamma_q(ik)}{\Gamma_q(k)},
\end{aligned} \tag{9.8}
$$

respectively, with l replaced by $n-m-l$. This establishes the equivalence of our formulas (5.6), (8.17), (9.7), and (9.8) for ${}_qI_n$, ${}_qJ_n$, ${}_qF_n$, and ${}_qG_n$, respectively.

Our entire analysis goes through as before with the cases $l=0$ and $m+l=n$ interchanged. Carlitz's q-analogue (5.18) of Dixon's theorem and its companion (5.17) also arise in the case $n=2$. Rather than give a direct proof, we may use (9.5) and (9.6) to "transport" Theorems 10 and 12 to apply to ${}_qF_n$ and ${}_qG_n$. For $k=1$, we may take $m+l=n$ and use (3.1) and (4.26) with $A=-c$.

Askey's last conjecture in [6] is based upon

$$
\int_{-c}^{d} \frac{(-tc^{-1}q)_\infty (td^{-1}q)_\infty}{(-tc^{-1}q^x)_\infty (td^{-1}q^y)_\infty} d_qt = \frac{cd}{(c+d)} \frac{(-cd^{-1})_\infty (-dc^{-1})_\infty}{(-q^ycd^{-1})_\infty (-q^xdc^{-1})_\infty} \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}, \tag{9.9}
$$

where there are no zero factors in the denominator of the integrand and

$$
\int_{-c}^{d} f(t)d_qt = c(1-q)\sum_{n=0}^{\infty} q^n f(-cq^n) + d(1-q)\sum_{n=0}^{\infty} q^n f(dq^n). \tag{9.10}
$$

This is given by Andrews and Askey [3]. We set

$$
{}_q^{c,d}H_{n,m,l}(x,y,k)=\int_{-c}^{d}\cdots\int_{-c}^{d}\prod_{i=1}^{n}\frac{(-t_ic^{-1}q)_\infty}{(-t_ic^{-1}q^{x+\chi(i\leqq m)})_\infty}\frac{(t_id^{-1}q)_\infty}{(t_id^{-1}q^{y+\chi(n-i+1\leqq l)})_\infty}
$$

$$
\cdot {}_{(q^k)}\Delta_{n,m,l}(\mathbf{t})\ {}_q\Delta_n^{(2k-1)}(\mathbf{t})\ d_qt_1\cdots d_qt_n \tag{9.11}
$$

and

$$
{}_q^{c,d}K_{n,m,l}(x,y,k)=\int_{-c}^{d}\cdots\int_{-c}^{d}\prod_{i=1}^{n}\frac{(-t_ic^{-1}q^{\chi(i>m)})_\infty}{(-t_ic^{-1}q^{x})_\infty}\frac{(t_id^{-1}q^{\chi(i\leqq n-l)})_\infty}{(t_id^{-1}q^{y})_\infty}
$$

$$
\cdot {}_{(q^{-k})}\Delta_{n,m,l}(\mathbf{t})\ {}_q\Delta_n^{(2k-1)}(\mathbf{t})\ d_qt_1\cdots d_qt_n. \tag{9.12}
$$

We shall work with ${}_qH_n$. We have the following conjecture.

Conjecture 13.

$$
\begin{aligned}
{}_q^{c,d}H_{n,m,l}(x,y,k)=n!&\frac{\begin{bmatrix}n\\m,\,l\end{bmatrix}_{(q^k)}}{\binom{n}{m,\,l}}q^{[k\binom{m}{2}+k\binom{l}{2}-\binom{k}{2}\binom{n}{2}+k^2\binom{n}{3}]}\\
&\cdot\prod_{i=1}^{n}\frac{(cd)^{1+(n-i)k}}{(c+d)}\frac{(-cd^{-1})_\infty}{(-q^{y+(n-i)k+\chi(i\leqq l)}cd^{-1})_\infty}\\
&\cdot\frac{(-dc^{-1})_\infty}{(-q^{x+(n-i)k+\chi(i\leqq m)}dc^{-1})_\infty}\\
&\cdot\prod_{i=1}^{n}\frac{\Gamma_q(x+(n-i)k+\chi(i\leqq m))\Gamma_q(y+(n-i)k+\chi(i\leqq l))}{\Gamma_q(x+y+(2n-i-1)k+\chi(i\leqq m+l))}\frac{\Gamma_q(ik)}{\Gamma_q(k)}.
\end{aligned} \tag{9.13}
$$

The substitution $t_i\to -t_{n-i+1}$, $1\leqq i\leqq n$, gives the q-analogue

$$
{}_q^{c,d}H_{n,m,l}(x,y,k)={}_q^{d,c}H_{n,l,m}(y,x,k) \tag{9.14}
$$

of the symmetry (1.3). We cannot use this substitution for ${}_qI_n$, ${}_qJ_n$, ${}_qF_n$, or ${}_qG_n$ since the lower limit of integration is 0. The substitution $t_i\to t_{n-i+1}^{-1}$, $1\leqq i\leqq n$, only gives again the equivalence of our formulas (9.7) and (9.8) for ${}_qF_n$ and ${}_qG_n$. Unfortunately, (9.14) does not imply the symmetry law (5.7) of ${}_qI_n$. If we could establish (5.7) (or an equivalent symmetry of ${}_qJ_n$, ${}_qF_n$, or ${}_qG_n$), then we should be able to extend Selberg's proof [27] by expanding our q-analogue of $\Delta_n^{2k}(\mathbf{t})$ in powers of t_i, $1\leqq i\leqq n$. We cannot make use of the symmetry (9.14) since we must expand in terms of $(-tc^{-1}q^x)_m$, $1\leqq i\leqq n$, and $(td^{-1}q^y)_m$, $1\leqq i\leqq n$. We can do this for $k=1$ using Lemma 5. Our proof of the case $k=1$ works because our key results (6.9) and (7.7) show that ${}_qU_{n,m}(x,y,1)$ and ${}_qU_{n,\lambda}(x,y,1)$ are symmetric in x and y. This gives the case $k=1$, $m=l=0$, of (5.7).

The usual argument for obtaining a recurrence relation gives

$$
\begin{aligned}
&\lim_{x\to 0}\frac{(1-q^x)}{(1-q)}\,{}_q^{c,d}H_{n,m}(x,y,k)\\
&\quad=(n-m)\frac{cd}{(c+d)}\frac{(-cd^{-1})_\infty}{(-q^ycd^{-1})_\infty}q^{km}c^{(n-m-1)}\\
&\qquad\cdot\int_{-c}^{d}\cdots\int_{-c}^{d}\prod_{i=1}^{n-1}t_i^{(2k-1)+\chi(i\leqq m)}\left(\frac{-cq^{1-k-\chi(i\leqq m)}}{t_i}\right)_{(2k-1)+\chi(i\leqq m)}\\
&\qquad\cdot\prod_{i=1}^{n-1}\frac{(t_id^{-1}q)_\infty}{(t_id^{-1}q^y)_\infty}\,{}_{(q^k)}\Delta_{n-1,m}(\mathbf{t})\ {}_q\Delta_{n-1}^{(2k-1)}(\mathbf{t})\ d_qt_1\cdots d_qt_{n-1}.
\end{aligned} \tag{9.15}
$$

Since

$$t_i^{(2k-1)+\chi(i\leqq m)}\left(\frac{-cq^{1-k-\chi(i\leqq m)}}{t_i}\right)_{(2k-1)+\chi(i\leqq m)} = c^{(2k-1)}(cq^{-k})^{\chi(i\leqq m)}(-t_iq^{1-k}c^{-1})_{(2k-1)+\chi(i\leqq m)}, \tag{9.16}$$

we obtain

$$\lim_{x\to 0}\frac{(1-q^x)}{(1-q)}\,{}_q^{c,d}H_{n,m}(x,y,k) = (n-m)\frac{cd}{(c+d)}\frac{(-cd^{-1})_\infty}{(-q^ycd^{-1})_\infty}c^{2k(n-1)}\,{}_q^{cq^k,d}H_{n-1,m}(2k,y,k). \tag{9.17}$$

To treat the case $k=1$, set $A=-q^xc^{-1}$ and $A=q^yd^{-1}$ in (4.26). In short order we have

$$\begin{aligned}{}_q^{c,d}H_{n,m}(x,y,1) &= m!(n-m)!q^{-[x\binom{n}{2}+y\binom{n}{2}+2\binom{n}{3}]}\\ &\cdot\prod_{i=1}^{n}\frac{(cd)^{1+n-i}}{(c+d)}\frac{(-cd^{-1})_\infty}{(-q^{y+n-i}cd^{-1})_\infty}\frac{(-dc^{-1})_\infty}{(-q^{x+n-i+\chi(i\leqq m)}dc^{-1})_\infty}\\ &\cdot\prod_{i=1}^{n}\Gamma_q(x+n-i+\chi(i\leqq m))\Gamma_q(y+n-i)\\ &\cdot\sum_{\pi\in S_n}\operatorname{sgn}(\pi)\prod_{i=1}^{n}\frac{1}{\Gamma_q(x+y+n-i+\pi(i)-1+\chi(i\leqq m))}.\end{aligned} \tag{9.18}$$

Observe that the sum on the right side of (9.18) equals ${}_qU_{n,m}(x,y,1)$ (6.9), which we have already (6.14) evaluated. Equation (9.13) follows for $k=1$, $l=0$.

Since (6.14) also arises in connection with orthogonal polynomials and plane partitions, it is an important sum. As with all of our q-analogues of Selberg's integral, it is really a polynomial identity. Set $z=x+y$ and multiply both sides of (6.14) by

$$(q-1)^{\binom{n}{2}}\prod_{i=1}^{n}\Gamma_q(z+2n-i-1+\chi(i\leqq m)). \tag{9.19}$$

Equation (6.14) then becomes

$$\begin{aligned}&\sum_{\pi\in S_n}\operatorname{sgn}(\pi)\prod_{i=1}^{n}(q^{z+2n-i-\pi(i)+\chi(i\leqq m)})_{(\pi(i)-1)}\\ &\quad=q^{[z\binom{n}{2}+2\binom{n}{2}+5\binom{n}{3}]}\prod_{1\leqq i<j\leqq n}(q^{-i+\chi(i\leqq m)}-q^{-j+\chi(j\leqq m)}),\end{aligned} \tag{9.20}$$

which is equivalent to the case (4.28) of Lemma 5 with $A=q^{z+n}$ and $t_j=q^{j-\chi(j\leqq m)}$, $1\leqq j\leqq n$.

We may give the same type of results for ${}_q^{c,d}K_{n,m,l}$. The formula is

$${}_q^{c,d}K_{n,m,l}(x,y,k) = q^{[xl+ym]}\,{}_q^{c,d}H_{n,m,l}(x,y,k). \tag{9.21}$$

An important generalization of the Schur function is given by

$${}_q^{A}S_{n,\lambda}(\mathbf{t}) = \frac{\det|(At_j)_{n-i+\lambda_i}|_{n\times n}}{\det|t_j^{n-i}|_{n\times n}}. \tag{9.22}$$

Set $A=-q^x c^{-1}$ in (9.22) and $A=q^y d^{-1}$ in (4.26). Using our evaluation (7.12) of ${}_qU_{n,\lambda}(x, y, 1)$, we obtain

$$
\begin{aligned}
{}_q^{c,d}I_{n,\lambda}(x, y, 1) &= \int_{-c}^{d} \cdots \int_{-c}^{d} \prod_{i=1}^{n} \frac{(-t_i c^{-1} q)_\infty}{(-t_i c^{-1} q^x)_\infty} \frac{(t_i d^{-1} q)_\infty}{(t_i d^{-1} q^y)_\infty} \\
&\quad \cdot {}^{(-q^x c^{-1})}_{q}S_{n,\lambda}(\mathbf{t})\, \Delta_n^2(\mathbf{t})\, d_q t_1 \cdots d_q t_n \\
&= n!\, q^{[\sum_{i=1}^{n}(i-1)\lambda_i + x\binom{n}{2} + 2\binom{n}{3}]} d^{\binom{n}{2}} \prod_{1 \le i < j \le n} \frac{(1-q^{j-i+\lambda_i-\lambda_j})}{(1-q)} \\
&\quad \cdot \prod_{i=1}^{n} \frac{cd}{(c+d)} \frac{(-cd^{-1})_\infty}{(-q^{y+n-i}cd^{-1})_\infty} \frac{(-dc^{-1})_\infty}{(-q^{x+n-i+\lambda_i}dc^{-1})_\infty} \\
&\quad \cdot \prod_{i=1}^{n} \frac{\Gamma_q(x+n-i+\lambda_i)\Gamma_q(y+n-i)}{\Gamma_q(x+y+2n-i-1+\lambda_i)}.
\end{aligned} \tag{9.23}
$$

Askey and Wilson [7] give the orthogonal polynomials for an important q-analogue of the beta distribution. Rahman [25] conjectures a q-analogue of Selberg's integral using this measure. It would be interesting to apply our methods in this case.

Note added in proof. Aomoto [A] has proven Conjecture 1 for $l=0$. The conjecture follows by (3.18) and (3.19). Kadell [K] extends Aomoto's proof to treat (1.11) and Conjectures 6 and 11.

[A] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.

[K] K. W. J. KADELL, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969-986.

REFERENCES

[1] G. E. ANDREWS, *Problems and prospects for basic hypergeometric functions*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 191-224.

[2] ———, *The Theory of Partitions, Encyclopedia of Mathematics and Its Applications*, 2, Addison-Wesley, Reading, MA, 1976.

[3] G. E. ANDREWS AND R. ASKEY, *Another q-extension of the beta function*, Proc. Amer. Math. Soc., 81 (1981), pp. 97-100.

[4] R. ASKEY, ed., *Theory and Application of Special Functions*, Academic Press, New York, 1975.

[5] ———, *The q-gamma and q-beta functions*, Applicable Anal., 8 (1978), pp. 125-141.

[6] ———, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.

[7] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Mem. Amer. Math. Soc. 319 (1985).

[8] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London, 1935; reprinted by Hafner, New York, 1964.

[9] L. CARLITZ, *Some formulas of F. H. Jackson*, Monatsh. Math., 73 (1969), pp. 193-198.

[10] R. W. CARTER, *Simple Groups of Lie Type*, Wiley-Interscience, London, New York, 1972.

[11] R. J. EVANS, *Identities for products of Gauss sums over finite fields*, Enseign. Math., 27 (1981), pp. 197-209.

[12a] B. GORDON AND L. HOUTEN, *Notes on plane partitions*, I, J. Combin. Theory, 4 (1968), pp. 72-80.

[12b] ———, *Notes on plane partitions*, II, J. Combin. Theory, 4 (1968), pp. 81-99.

[13] F. H. JACKSON, *On q-definite integrals*, Quart. J. Pure Appl. Math., 41 (1910), pp. 193-203.

[14] ———, *Certain q-identities*, Quart. J. Math., 12 (1941), pp. 167-172.

[15] K. W. J. KADELL, *Weighted inversion numbers, restricted growth functions, and standard Young tableaux*, J. Combin. Theory Ser. A, 40 (1985), pp. 22-44.

[16] ———, *Andrews' q-Dyson conjecture* II: *Symmetry*, Pacific J. Math., to appear.

[17] I. G. MACDONALD, *The Poincaré series of a Coxeter group*, Math. Ann., 199 (1972), pp. 161-174.

[18] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Clarendon Press, Oxford, 1979.
[19] ———, Private communication to Richard Askey, June 1980.
[20] ———, *Some conjectures for root systems and finite reflection groups*, SIAM J. Math. Anal., 13 (1982), pp. 988-1007.
[21] P. A. MACMAHON, *Two applications of general theorems in combinatory analysis*: (1) *to the theory of inversions of permutations*; (2) *to the ascertainment of the numbers of terms in the development of a determinant which has amongst its elements an arbitrary number of zeros*, Proc. London Math. Soc. (2), 15 (1916), pp. 314-321.
[22] M. L. MEHTA, *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York, 1967.
[23] R. A. MENA, *On discriminant-like polynomials*, Linear Algebra Appl., to appear.
[24] W. MORRIS, *Constant term identities for finite and affine root systems*, Ph.D. thesis, Univ. of Wisconsin, Madison, WI, 1982.
[25] M. RAHMAN, *Another conjectured q-Selberg integral*, SIAM J. Math. Anal., 17 (1986), pp. 1267-1279.
[26] D. ST. P. RICHARDS, *Integrals of Selberg polynomials*, preprint.
[27] A. SELBERG, *Bemerkninger om et multipelt integral*, Nordisk Tidskr., 26 (1944), pp. 71-78.
[28] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, London, 1966.
[29] D. ZEILBERGER AND D. M. BRESSOUD, *A proof of Andrews' q-Dyson conjecture*, Discrete Math., 54 (1985), pp. 201-224.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

015

A PROOF OF ASKEY'S CONJECTURED q-ANALOGUE OF SELBERG'S INTEGRAL AND A CONJECTURE OF MORRIS*

KEVIN W. J. KADELL†

Abstract. Aomoto has recently given an extension of Selberg's integral. We extend his proof to the q-case and establish a conjecture of Askey. Our result is equivalent to a constant term identity for the root system A_n. This extends a conjecture of Morris.

Key words. Selberg's integral, Aomoto's extension, q-beta integral, Morris' theorem, Cauchy-Selberg labeling for A_n

AMS(MOS) subject classification. 33A15

1. Introduction and summary. In 1944, Selberg [15] gave an elegant evaluation of an important multivariable beta type integral. Macdonald [12] obtained the case $q=1$ of his constant term conjectures for B_n, C_n, D_n and BC_n. Recently, Aomoto [4] proved an extension of Selberg's integral to which we may add the parameter l. This is the following.

THEOREM 1 (Aomoto [4]).

$$I_{n,m,l}(x,y,k)=\int_0^1\cdots\int_0^1\prod_{i=1}^{n}t_i^{(x-1)+\chi(i\leqq m)}(1-t_i)^{(y-1)+\chi(n-i+1\leqq l)}\Delta_n^{2k}(\mathbf{t})\,dt_1\cdots dt_n$$

$$=\prod_{i=1}^{n}\frac{\Gamma(x+(n-i)k+\chi(i\leqq m))\Gamma(y+(n-i)k+\chi(i\leqq l))\Gamma(1+ik)}{\Gamma(x+y+(2n-i-1)k+\chi(i\leqq m+l))\Gamma(1+k)}, \tag{1.1}$$

where $\chi(A)$ is 1 or 0 according to whether A is true or false,

$$\Delta_n(t_1,\cdots,t_n)=\Delta_n(\mathbf{t})=\prod_{1\leqq i<j\leqq n}(t_i-t_j), \tag{1.2}$$

and, as holds throughout, $\operatorname{Re}(x)>0$, $\operatorname{Re}(y)>0$, *and n, m, l, and k are nonnegative integers satisfying $m+l\leqq n$. We omit l when $l=0$.*

The substitution $t_i\to(1-t_{n-i+1})$, $1\leqq i\leqq n$, gives the symmetry

$$I_{n,m,l}(x,y,k)=I_{n,l,m}(y,x,k). \tag{1.3}$$

This is essential to Selberg's proof [15] of the case $m=l=0$ of (1.1). See Andrews [3] for a readily accessible version of this argument. This proof did not work in the q-case since we could not give an appropriate symmetry for Askey's Conjecture 1 of [6]. Where the symmetry was restored [6, Conjecture 8], the argument failed for other reasons.

Aomoto's proof [4] of the case $l=0$ of Theorem 1 relies on a difference equation involving the extra parameter m. We extend this argument to treat a q-analogue of (1.1). Our result is Theorem 2.

* Received by the editors March 5, 1986; accepted for publication May 28, 1987. This research was supported by the Natural Sciences and Engineering Research Council of Canada under grants A8907 and A8235.

† Department of Mathematics, Arizona State University, Tempe, Arizona 85287. Present address, School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455. This research was done while the author was on a visiting appointment to the Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 during 1985-86.

THEOREM 2.

$$
\begin{aligned}
{}_qS_{n,m,l}(x,y,k) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i\leqq m)} \frac{(qt_i)_\infty}{(q^{y+\chi(n-i+1\leqq l)}t_i)_\infty} \\
&\quad \cdot \prod_{1\leqq i<j\leqq n} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k} d_qt_1\cdots d_qt_n \qquad (1.4)\\
&= q^{[kx\binom{n}{2}+k\binom{m}{2}+2k^2\binom{n}{3}]} \\
&\quad \cdot \prod_{i=1}^{n} \frac{\Gamma_q(x+(n-i)k+\chi(i\leqq m))\Gamma_q(y+(n-i)k+\chi(i\leqq l))\Gamma_q(1+ik)}{\Gamma_q(x+y+(2n-i-1)k+\chi(i\leqq m+l))\Gamma_q(1+k)},
\end{aligned}
$$

where q is fixed with $0<q<1$,

$$
\begin{aligned}
(x)_0 &= (x;q)_0 = 1, \\
(x)_n &= (x;q)_n = \prod_{i=0}^{n-1}(1-xq^i), \qquad n\geqq 1, \qquad (1.5)\\
(x)_\infty &= (x;q)_\infty = \lim_{n\to\infty}(x)_n = \prod_{i=0}^{\infty}(1-xq^i),
\end{aligned}
$$

and, following Jackson [8],

$$
\int_0^a f(t)\,d_qt = a(1-q)\sum_{n=0}^{\infty} q^n f(aq^n), \qquad (1.6)
$$

$$
\Gamma_q(x) = (1-q)^{(1-x)}\frac{(q)_\infty}{(q^x)_\infty}. \qquad (1.7)
$$

The key to Aomoto's proof [4] is the observation that if

$$
F(0,t_2,\cdots,t_n) = F(1,t_2,\cdots,t_n) = 0 \qquad (1.8)
$$

and F satisfies some simple conditions, then

$$
\int_0^1 \cdots \int_0^1 \frac{\partial}{\partial t_1} F(\mathbf{t})\,dt_1\cdots dt_n = 0. \qquad (1.9)
$$

Aomoto sets $l=0$ and takes F to be t_1^δ times the integrand in (1.1), where δ equals 0 or 1. Two simple lemmas allow us to eliminate the results of (1.9) for $\delta=0$ and $\delta=1$. This gives

$$
I_{n,m}(x,y,k) = \frac{(x+(n-m)k)}{(x+y+(2n-m-1)k)} I_{n,m-1}(x,y,k), \qquad m\geqq 1. \qquad (1.10)
$$

The case $l=0$ follows using a recurrence relation given by Selberg [15]. See also Kadell [10]. Since $1=t_i+(1-t_i)$, we have

$$
I_{n,m,l}(x,y,k) = I_{n,m+1,l}(x,y,k) + I_{n,m,l+1}(x,y,k). \qquad (1.11)
$$

Theorem 1 now follows by induction on l.

In § 2, we introduce the q-derivative corresponding to the q-integral (1.6). A simple telescoping sum provides a q-analogue of the fundamental theorem of calculus. We extend (1.9), thus taking the first step in the proof of Theorem 2.

In § 3, we give two lemmas that extend the basic steps of Aomoto's proof [4] to the q-case. Unfortunately, we cannot eliminate from our equations as easily as we can when $q=1$. This is because the integrand of (1.4) with $m=l=0$ is not symmetric in $t_1, t_2, \cdots, t_n$. In § 4, we recall Lemma 4 of Kadell [10] and give a useful alternative formulation. This allows us to treat all of the difficulties arising from the asymmetry. We also recall a recurrence relation [10, (6.4)] and obtain a recurrence relation for ${}_qS_{n,m}(x, y, k)$.

In § 5, we use simple manipulations of our results to obtain a q-analogue of (1.10). We then establish the case $l=0$ of Theorem 2.

By Lemma 4 of [10], this gives the case $l=0$ of Conjectures 6 and 11 of [10]. These conjectures now follow by Theorems 10 and 12 of [10]. In § 6, we use the same approach to prove Theorem 2.

In § 7, we show that Theorem 2 is equivalent to a constant term identity for the root system A_n. Set

$$
\begin{aligned}
{}_qcs_{n,m,l}(a, b, k; \mathbf{t}) = \prod_{i=1}^{n} (q^{\chi(i>m)}t_i)_{a+\chi(i\leqq m)+\chi(n-i+1\leqq l)}\left(\frac{q^{\chi(i\leqq m)}}{t_i}\right)_{b-\chi(i\leqq m)} \\
\cdot \prod_{1\leqq i<j\leqq n}\left(q\frac{t_j}{t_i}\right)_k\left(\frac{t_i}{t_j}\right)_k,
\end{aligned} \tag{1.12}
$$

where a and b are nonnegative integers. Let

$$
{}_qCS_{n,m,l}(a, b, k) = \text{C.T. } {}_qcs_{n,m,l}(a, b, k; \mathbf{t}), \tag{1.13}
$$

where C.T.$f(\mathbf{t})$ is the constant term in the Laurent expansion of $f(\mathbf{t})$ in powers of $t_1, t_2, \cdots, t_n$. Our result is

THEOREM 3.

$$
{}_qCS_{n,m,l}(a, b, k) = \prod_{i=1}^{n} \frac{(q)_{a+b+(n-i)k+\chi(i\leqq l)}(q)_{ik}}{(q)_{a+(n-i)k+\chi(i\leqq m+l)}(q)_{b+(i-1)k-\chi(i\leqq m)}(q)_k}. \tag{1.14}
$$

The case $m=l=0$ was conjectured by Morris [14, (4.12)]. He proved [14, § 6] the case $m=l=0$, $q=1$, by using a version of Selberg's integral which extends Cauchy's form of the beta integral.

2. q-derivatives. Set

$$
\frac{d_q}{d_qt}F(t) = \frac{F(t)-F(qt)}{t(1-q)}. \tag{2.1}
$$

A q-analogue of the fundamental theorem of calculus is given by

$$
\begin{aligned}
\int_0^a \frac{d_q}{d_qt}F(t)\, d_qt &= (1-q)\sum_{n=0}^{\infty} aq^n \frac{F(aq^n)-F(aq^{n+1})}{aq^n(1-q)} \\
&= \sum_{n=0}^{\infty} F(aq^n)-F(aq^{n+1}) = F(a)-\lim_{n\to\infty} F(aq^n).
\end{aligned} \tag{2.2}
$$

Let

$${}_qW_n(x, y, k) = \prod_{i=1}^{n} t_i^{(x-1)}\frac{(qt_i)_\infty}{(q^yt_i)_\infty}\prod_{1\leqq i<j\leqq n} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k} \tag{2.3}$$

denote the integrand in (1.4) when $m=l=0$. For $1\leqq i<j\leqq n$, the function

$${}_q\Delta_n^{(2k-1)}(\mathbf{t}) = \prod_{1\leqq i<j\leqq n} t_i^{(2k-1)}\left(q^{1-k}\frac{t_j}{t_i}\right)_{(2k-1)} \tag{2.4}$$

vanishes when $t_i = t_j$ and on $k-1$ lines on one side of this line and $k-1$ lines symmetrically located on the other side. Thus (see Kadell [10, (4.15)])

$$
\begin{aligned}
&{}_q\Delta_n^{(2k-1)}(\pi(\mathbf{t})) = \operatorname{sgn}(\pi)\, {}_q\Delta_n^{(2k-1)}(\mathbf{t}), \\
&\pi(\mathbf{t}) = (t_{\pi(1)}, t_{\pi(2)}, \cdots, t_{\pi(n)}),
\end{aligned} \tag{2.5}
$$

where $\pi \in S_n$. To obtain a q-analogue of $\Delta_n^{2k}(\mathbf{t})$, we may multiply ${}_q\Delta_n^{(2k-1)}(\mathbf{t})$ by $\prod_{1\leqq i<j\leqq n}(t_i - q^k t_j)$ (as in ${}_qW_n(x, y, k)$) or by $\prod_{1\leqq i<j\leqq n}(t_i - q^{-k}t_j)$. Thus ${}_qW_n(x, y, k)$ is not symmetric in $t_1, t_2, \cdots, t_n$ even though ${}_q\Delta_n^{(2k-1)}(\mathbf{t})$ is antisymmetric (2.5). The reader should carefully observe the role played by this last factor throughout the proof of Theorem 2.

We now assume that

$$
1 \leqq m \leqq n, \quad \operatorname{Re}(x) > 0, \quad \operatorname{Re}(y) > 1 \quad \text{and} \quad \delta \text{ is a nonnegative integer.} \tag{2.6}
$$

This assures that ${}_qW_n(x, y, k)$ is 0 if $t_i = 1/q$ for some i, $1 \leqq i \leqq n$. Hence, we may extend the range of each integration in (1.4) to $1/q$. We adopt the obvious notation for partial q-derivatives. Since $m \geqq 1$, we see that $\prod_{i=1}^m t_i$ vanishes when $t_1 = 0$. Using (2.2), we obtain

$$
0 = \int_0^{1/q} \cdots \int_0^{1/q} \frac{\partial_q}{\partial_q t_1}\left(t_1^{\delta} \prod_{i=1}^{m} t_i\, {}_qW_n(x, y, k)\right) d_q t_1 \cdots d_q t_n. \tag{2.7}
$$

In order to compute the partial q-derivative above, we require a product rule for q-derivatives. It is

$$
\begin{aligned}
\frac{d_q}{d_q t} \prod_{v=1}^{u} F_v(t) &= \frac{1}{t(1-q)}\left(\prod_{v=1}^{u} F_v(t) - \prod_{v=1}^{u} F_v(qt)\right) \\
&= \frac{1}{t(1-q)} \sum_{v=1}^{u} \prod_{i=1}^{v-1} F_i(qt)\left(F_v(t) - F_v(qt)\right) \prod_{j=v+1}^{u} F_j(t) \\
&= \sum_{v=1}^{u} \prod_{i=1}^{v-1} F_i(qt) \frac{d_q}{d_q t} F_v(t) \prod_{j=v+1}^{u} F_j(t).
\end{aligned} \tag{2.8}
$$

A few simple computations give

$$
\begin{aligned}
\frac{\partial_q}{\partial_q s}\left(s^{2k}\left(q^{1-k}\frac{t}{s}\right)_{2k}\right) &= \frac{(1-q^{2k})}{(1-q)} s^{(2k-1)}\left(q^{1-k}\frac{t}{s}\right)_{(2k-1)} \\
&= \frac{(1-q^{2k})}{(1-q)} \frac{1}{(s-q^k t)} s^{2k}\left(q^{1-k}\frac{t}{s}\right)_{2k},
\end{aligned} \tag{2.9}
$$

$$
\frac{d_q}{d_q t} t^x = \frac{(1-q^x)}{(1-q)} t^{(x-1)}, \tag{2.10}
$$

$$
\frac{d_q}{d_q t} \frac{(qt)_\infty}{(q^y t)_\infty} = -q\frac{(1-q^{(y-1)})}{(1-q)} \frac{1}{(1-qt)} \frac{(qt)_\infty}{(q^y t)_\infty}. \tag{2.11}
$$

Let v run from 2 to n and take $s = t_1$, $t = t_v$, in (2.4). Using our product rule (2.8) and

$$
(qs)^{2k}\left(q^{1-k}\frac{t}{qs}\right)_{2k} = q^{2k}\frac{(s-q^{-k}t)}{(s-q^k t)} s^{2k}\left(q^{1-k}\frac{t}{s}\right)_{2k}, \tag{2.12}
$$

we obtain

$$\frac{\partial_q}{\partial_q t_1}\left(\prod_{1\leqq i<j\leqq n} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k}\right)=\frac{(1-q^{2k})}{(1-q)}\sum_{v=2}^{n} q^{2k(v-2)}\prod_{j=2}^{v-1}\frac{(t_1-q^{-k}t_j)}{(t_1-q^k t_j)} \cdot\frac{1}{(t_1-q^k t_v)}\prod_{1\leqq i<j\leqq n} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k}. \tag{2.13}$$

Continuing with our product rule (2.8), we use (2.10) followed by (2.11) to compute the partial q-derivative in (2.7). We obtain

$$0=\frac{(1-q^{2k})}{(1-q)}\sum_{v=2}^{n} q^{2k(v-2)}\,{}_q^{\delta}A_{n,m}^{v}(x,y,k)+q^{2k(n-1)}\frac{(1-q^{x+\delta})}{(1-q)}\,{}_q^{\delta}K_{n,m}(x,y,k) -q^{2k(n-1)+x+\delta+1}\frac{(1-q^{(y-1)})}{(1-q)}\,{}_q^{\delta}E_{n,m}(x,y,k), \tag{2.14}$$

where

$${}_q^{\delta}A_{n,m}^{v}(x,y,k)=\int_0^{1/q}\cdots\int_0^{1/q}\prod_{j=2}^{v-1}\frac{(t_1-q^{-k}t_j)}{(t_1-q^k t_j)}\frac{1}{(t_1-q^k t_v)} \cdot t_1^{\delta}\prod_{i=1}^{m} t_i\,{}_qW_n(x,y,k)\,d_qt_1\cdots d_qt_n, \tag{2.15}$$

$${}_q^{\delta}K_{n,m}(x,y,k)=\int_0^{1/q}\cdots\int_0^{1/q}\prod_{j=2}^{n}\frac{(t_1-q^{-k}t_j)}{(t_1-q^k t_j)} \cdot t_1^{\delta}\prod_{i=2}^{m} t_i\,{}_qW_n(x,y,k)\,d_qt_1\cdots d_qt_n, \tag{2.16}$$

$${}_q^{\delta}E_{n,m}(x,y,k)=\int_0^{1/q}\cdots\int_0^{1/q}\prod_{j=2}^{n}\frac{(t_1-q^{-k}t_j)}{(t_1-q^k t_j)}\frac{1}{(1-qt_1)} \cdot t_1^{\delta}\prod_{i=1}^{m} t_i\,{}_qW_n(x,y,k)\,d_qt_1\cdots d_qt_n. \tag{2.17}$$

3. Two lemmas. The effect of the factor $(t_1-q^{-k}t_j)/(t_1-q^k t_j)$ in (2.15), (2.16),and (2.17) is to replace the factor $(t_1-q^k t_j)$ of ${}_qW_n(x,y,k)$ (which "glues onto" one end of ${}_q\Delta_n^{(2k-1)}(\mathbf{t})$) by $(t_1-q^{-k}t_j)$ (which "glues onto" the other end). Observe that

$$\prod_{j=2}^{v-1}\frac{(t_1-q^{-k}t_j)}{(t_1-q^k t_j)}\frac{1}{(t_1-q^k t_v)}\,{}_qW_n(x,y,k)=\prod_{i=1}^{n} t_i^{(x-1)}\frac{(qt_i)_\infty}{(q^y t_i)_\infty}\,{}_q\Delta_n^{(2k-1)}(\mathbf{t}) \cdot\prod_{\substack{2\leqq i<j\leqq n\\ i\neq v\neq j}}(t_i-q^k t_j)\prod_{j=v+1}^{n}(t_1-q^k t_j)(t_v-q^k t_j) \cdot\prod_{j=2}^{v-1}(t_1-q^{-k}t_j)\prod_{i=2}^{v-1}(t_i-q^k t_v). \tag{3.1}$$

Since

$$(t_1-q^{-k}t_j)=-q^{-k}(t_j-q^k t_1), \tag{3.2}$$

the function (3.1) is antisymmetric in t_1 and t_v. We have Lemma 4.

LEMMA 4. *Let a be a nonnegative integer and* Sym (t) *be symmetric in t_1 and t_v where $2 \leqq v \leqq n$. Then*

$$\int_0^{1/q} \cdots \int_0^{1/q} t_1^a \operatorname{Sym}(\mathbf{t}) \prod_{j=2}^{v-1} \frac{(t_1 - q^{-k}t_j)}{(t_1 - q^k t_j)} \frac{1}{(t_1 - q^k t_v)} {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n$$

$$= \frac{1}{(1+q^{ak})} \int_0^{1/q} \cdots \int_0^{1/q} \left(\sum_{i=0}^{a-1} t_1^i (q^k t_v)^{(a-1-i)} \right) \operatorname{Sym}(\mathbf{t}) \tag{3.3}$$

$$\cdot \prod_{j=2}^{v-1} \frac{(t_1 - q^{-k}t_j)}{(t_1 - q^k t_j)} {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n.$$

Proof. Let X denote the integral in (3.3). Interchanging t_1 and t_v gives

$$X = -\int_0^{1/q} \cdots \int_0^{1/q} t_v^a \operatorname{Sym}(\mathbf{t}) \prod_{j=2}^{v-1} \frac{(t_1 - q^{-k}t_j)}{(t_1 - q^k t_j)} \frac{1}{(t_1 - q^k t_v)} {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n, \tag{3.4}$$

since Sym (t) and (3.1) are symmetric and antisymmetric, respectively, in t_1 and t_v. Observe for $a = 0$ that $X = -X$ and hence $X = 0$. We have

$$\frac{(t_1^a - q^{ak} t_v^a)}{(t_1 - q^k t_v)} = \sum_{i=0}^{a-1} t_1^i (q^k t_v)^{(a-1-i)}. \tag{3.5}$$

The result (3.3) now follows by adding q^{ak} times (3.4) to X (3.3), simplifying the integrand by (3.5), and dividing by $(1+q^{ak})$. □

We now apply this lemma to ${}^{\delta}_{q}A^{v}_{n,m}(x, y, k)$ where δ equals 0 or 1. For $2 \leqq v \leqq m$, we may take $a = \delta$, Sym (t) $= \prod_{i=1}^{m} t_i$. This gives

$${}^{0}_{q}A^{v}_{n,m}(x, y, k) = 0, \qquad 2 \leqq v \leqq m, \tag{3.6}$$

$${}^{1}_{q}A^{v}_{n,m}(x, y, k) = \frac{1}{(1+q^k)} \int_0^{1/q} \cdots \int_0^{1/q} \prod_{i=1}^{m} t_i \prod_{j=2}^{v-1} \frac{(t_1 - q^{-k}t_j)}{(t_1 - q^k t_j)}$$

$$\cdot {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n, \qquad 2 \leqq v \leqq m. \tag{3.7}$$

For $m < v \leqq n$, we may take $a = \delta + 1$, Sym (t) $= \prod_{i=2}^{m} t_i$. We obtain

$${}^{0}_{q}A^{v}_{n,m}(x, y, k) = \frac{1}{(1+q^k)} \int_0^{1/q} \cdots \int_0^{1/q} \prod_{i=2}^{m} t_i \prod_{j=2}^{v-1} \frac{(t_1 - q^{-k}t_j)}{(t_1 - q^k t_j)}$$

$$\cdot {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n, \qquad m < v \leqq n, \tag{3.8}$$

$${}^{1}_{q}A^{v}_{n,m}(x, y, k) = \frac{1}{(1+q^{2k})} \int_0^{1/q} \cdots \int_0^{1/q} \prod_{i=2}^{m} t_i (t_1 + q^k t_v) \prod_{j=2}^{v-1} \frac{(t_1 - q^{-k}t_j)}{(t_1 - q^k t_j)}$$

$$\cdot {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n, \qquad m < v \leqq n. \tag{3.9}$$

We have a simple lemma.

LEMMA 5.

$${}^{\delta}_{q}E_{n,m}(x, y, k) = {}^{\delta+1}_{q}K_{n,m}(x, y, k) + q\, {}^{\delta+1}_{q}E_{n,m}(x, y, k). \tag{3.10}$$

Proof. The integrand (2.17) of ${}^{\delta}_{q}E_{n,m}(x, y, k)$ equals

$$\frac{t_1}{(1 - qt_1)} = t_1 + \frac{qt_1^2}{(1 - qt_1)} \tag{3.11}$$

times the integrand (2.16) of ${}^{\delta}_{q}K_{n,m}(x, y, k)$. The result (3.10) follows directly by using (3.11) to expand the integrand (2.17) of ${}^{\delta}_{q}E_{n,m}(x, y, k)$. □

4. A basic lemma. Lemma 4 of Kadell [10] enables us to treat factors of the form $(t_1-q^{-k}t_j)/(t_1-q^k t_j)$. We require the integral formulation [10, (4.9)]. This is

$$\int\cdots\int \prod_{i=1}^{m} t_i \prod_{1\leqq i<j\leqq n}(t_i-Q_{i,j}t_j)\,d_\mu(\mathbf{t}) = \left[\sum_{\pi\in S_{n,m}} \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j}\right] \cdot \int\cdots\int \prod_{i=1}^{n} t_i^{(n-i)+\chi(i\leqq m)}\,d_\mu(\mathbf{t}), \tag{4.1}$$

where $d_\mu(\mathbf{t})$ is an n-dimensional antisymmetric measure and

$$S_{n,m}=\{\pi\in S_n \mid \pi(i)\leqq m \text{ iff } i\leqq m\}. \tag{4.2}$$

For any choice of $Q_{i,j}$, $1\leqq i<j\leqq n$, we need only be concerned with the sum in brackets on the right side of (4.1). Since

$$(t_i-Q_{i,j}t_j)=-Q_{i,j}(t_j-Q_{i,j}^{-1}t_i), \tag{4.3}$$

we may permute the t_i, $1\leqq i\leqq n$, thus changing $\prod_{i=1}^m t_i$ to the product of any m of the t_i, $1\leqq i\leqq n$. This is precisely what we require and, fortunately, there is an easy way to do the computation.

In the proof of Lemma 4 of [10], we showed the following. The terms in the expansion of $\prod_{i=1}^m t_i \prod_{1\leqq i<j\leqq n}(t_i-Q_{i,j}t_j)$ in which no two of the variables t_i, $1\leqq i\leqq n$, occur to the same power are of the form

$$\prod_{i=1}^{m} t_i \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)<\pi(j)}} t_i \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}}(-Q_{i,j}t_j)=\operatorname{sgn}(\pi) \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j} \prod_{i=1}^{n} t_i^{(n-\pi(i))+\chi(i\leqq m)}, \tag{4.4}$$

where $\pi\in S_{n,m}$. Since $d_\mu(\mathbf{t})$ is antisymmetric (2.5), the other terms contribute 0 to the integral (4.1). The substitution $t_i\to t_{\pi(i)}$, $1\leqq i\leqq n$, gives $\prod_{1\leqq i<j\leqq n,\pi(i)>\pi(j)} Q_{i,j}$ times the integral on the right side of (4.1).

Let $[s,t]=\{i\in Z \mid s\leqq i\leqq t\}$ and let $M\subset[1,n]$ with $|M|=m$. The terms in the expansion of $\prod_{i\in M} t_i \prod_{1\leqq i<j\leqq n}(t_i-Q_{i,j}t_j)$ with distinct exponents are of the form

$$\prod_{i\in M} t_i \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)<\pi(j)}} t_i \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}}(-Q_{i,j}t_j)=\operatorname{sgn}(\pi) \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j} \prod_{i=1}^{n} t_i^{(n-\pi(i))+\chi(i\in M)}, \tag{4.5}$$

where π is in

$$S_{n,M}=\{\pi\in S_n \mid \pi(i)\leqq m \text{ iff } i\in M\}. \tag{4.6}$$

We obtain

$$\int\cdots\int \prod_{i\in M} t_i \prod_{1\leqq i<j\leqq n}(t_i-Q_{i,j}t_j)\,d_\mu(\mathbf{t}) = \left[\sum_{\pi\in S_{n,M}} \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q_{i,j}\right] \cdot \int\cdots\int \prod_{i=1}^{n} t_i^{(n-i)+\chi(i\leqq m)}\,d_\mu(\mathbf{t}). \tag{4.7}$$

It is well known (see MacMahon [13]) that

$$\sum_{\pi\in S_n} \prod_{\substack{1\leqq i<j\leqq n\\ \pi(i)>\pi(j)}} Q=\frac{(Q;Q)_n}{(1-Q)^n}. \tag{4.8}$$

Q marks the inversions $1\leqq i<j\leqq n$, $\pi(i)>\pi(j)$, of π. We may view π as a word $\pi(i)$, $1\leqq i\leqq n$, and let the letters be any set of n distinct numbers.

Let $\pi \in S_{n,M}$ and consider i, j with $1 \leqq i < j \leqq n$. If $i \in M$, $j \notin M$, then $\pi(i) \leqq m < \pi(j)$. If $i \notin M$, $j \in M$, then $\pi(i) > m \geqq \pi(j)$. This gives

$$e(M) = \sum_{\substack{1 \leqq i < j \leqq n \\ i \notin M, j \in M}} 1 \tag{4.9}$$

inversions of π. The remaining inversions have either $i, j \in M$ or $i, j \notin M$. These arise independently from the subwords $\pi(i)$, $i \in M$ and $\pi(i)$, $i \in [1, n] - M$, respectively. See Kendall and Stuart [11, pp. 496-512] and Goulden and Jackson [7, pp. 96-99] for the multinomial version of this argument. We obtain

$$\sum_{\pi \in S_{n,M}} \prod_{\substack{1 \leqq i < j \leqq n \\ \pi(i) > \pi(j)}} Q = Q^{e(M)} \frac{(Q; Q)_m (Q; Q)_{n-m}}{(1-Q)^n}. \tag{4.10}$$

Equation (4.7) then becomes

$$\begin{aligned} \int \cdots \int \prod_{i \in M} t_i \prod_{1 \leqq i < j \leqq n} (t_i - Qt_j)\, d_\mu(\mathbf{t}) &= Q^{e(M)} \frac{(Q; Q)_m (Q; Q)_{n-m}}{(1-Q)^n} \\ &\quad \cdot \int \cdots \int \prod_{i=1}^{n} t_i^{(n-i)+\chi(i \leqq m)}\, d_\mu(\mathbf{t}). \end{aligned} \tag{4.11}$$

Observe that the integral on the right side of (4.11) is independent of Q and M. We have

$$e([1, m]) = 0, \qquad \operatorname{card}(S_{n,M}) = m!(n-m)!. \tag{4.12}$$

Set $M = [1, m]$ and compare (4.11) and the case $Q = 1$. This yields

$$\begin{aligned} \int \cdots \int \prod_{i=1}^{m} t_i \prod_{1 \leqq i < j \leqq n} (t_i - Qt_j)\, d_\mu(\mathbf{t}) &= \frac{(Q; Q)_m (Q; Q)_{n-m}}{(1-Q)^n} \frac{1}{m!(n-m)!} \\ &\quad \cdot \int \cdots \int \prod_{i=1}^{m} t_i \prod_{1 \leqq i < j \leqq n} (t_i - t_j)\, d_\mu(\mathbf{t}). \end{aligned} \tag{4.13}$$

Comparing (4.11) and the case $M = [1, m]$, we obtain

$$\begin{aligned} &\int \cdots \int \prod_{i \in M} t_i \prod_{1 \leqq i < j \leqq n} (t_i - Qt_j)\, d_\mu(\mathbf{t}) \\ &\quad = Q^{e(M)} \int \cdots \int \prod_{i=1}^{m} t_i \prod_{1 \leqq i < j \leqq n} (t_i - Qt_j)\, d_\mu(\mathbf{t}). \end{aligned} \tag{4.14}$$

By (2.5), the measure

$$d_\mu(\mathbf{t}) = \prod_{i=1}^{n} t_i^{(x-1)} \frac{(qt_i)_\infty}{(q^y t_i)_\infty}\, {}_q\Delta_n^{(2k-1)}(\mathbf{t})\, d_q t_1 \cdots d_q t_n \tag{4.15}$$

is antisymmetric. For $l = 0$, the integrand (1.4) of ${}_qS_{n,m}(x, y, k)$ is

$$\prod_{i=1}^{m} t_i\, {}_qW_n(x, y, k)\, d_q t_1 \cdots d_q t_n = \prod_{i=1}^{m} t_i \prod_{1 \leqq i < j \leqq n} (t_i - q^k t_j)\, d_\mu(\mathbf{t}). \tag{4.16}$$

In [10], we studied

$$\begin{aligned} {}_qI_{n,m,l}(x, y, k) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i \leqq m)} \frac{(qt_i)_\infty}{(q^{y+\chi(n-i+1 \leqq l)} t_i)_\infty} \\ &\quad \cdot \Delta_n(t_1, \cdots, t_{n-l}, q^k t_{n-l+1}, \cdots, q^k t_n) \\ &\quad \cdot {}_q\Delta_n^{(2k-1)}(\mathbf{t})\, d_q t_1 \cdots d_q t_n. \end{aligned} \tag{4.17}$$

For $l=0$, the integrand is

$$\prod_{i=1}^{m} t_i \prod_{1\leqq i<j\leqq n} (t_i - t_j)\, d_\mu(\mathbf{t}). \tag{4.18}$$

For $Q=q^k$, (4.13) becomes

$$ {}_qS_{n,m}(x, y, k) = \frac{(q^k; q^k)_m (q^k; q^k)_{n-m}}{(1-q^k)^n} \frac{1}{m!(n-m)!}\, {}_qI_{n,m}(x, y, k). \tag{4.19}$$

We have the recurrence relation [10, (6.4)]

$$\lim_{x\to 0} \frac{(1-q^x)}{(1-q)}\, {}_qI_{n,m}(x, y, k) = (n-m)\, {}_qI_{n-1,m}(2k, y, k). \tag{4.20}$$

Using (4.19) or the same analysis, we obtain

$$\lim_{x\to 0} \frac{(1-q^x)}{(1-q)}\, {}_qS_{n,m}(x, y, k) = \frac{(1-q^{k(n-m)})}{(1-q^k)}\, {}_qS_{n-1,m}(2k, y, k). \tag{4.21}$$

Let $1\leqq s\leqq n$. The permutations

$$\begin{aligned} \pi_s(\mathbf{t}) &= (t_s, t_1, \cdots, t_{s-1}, t_{s+1}, \cdots, t_n), \\ \sigma_s(\mathbf{t}) &= (t_2, \cdots, t_s, t_1, t_{s+1}, \cdots, t_n), \end{aligned} \tag{4.22}$$

are inverses and

$$\operatorname{sgn}(\pi_s) = \operatorname{sgn}(\sigma_s) = (-1)^{(s-1)}. \tag{4.23}$$

Observe that

$$\begin{aligned} \prod_{j=2}^{s} \frac{(t_1 - Q^{-1}t_j)}{(t_1 - Qt_j)} \prod_{1\leqq i<j\leqq n} (t_i - Qt_j) &= \prod_{2\leqq i<j\leqq n} (t_i - Qt_j) \prod_{j=2}^{s} (t_1 - Q^{-1}t_j) \prod_{j=s+1}^{n} (t_1 - Qt_j) \\ &= (-Q^{-1})^{(s-1)} \prod_{2\leqq i<j\leqq n} (t_i - Qt_j) \prod_{j=2}^{s} (t_j - Qt_1) \\ &\quad \cdot \prod_{j=s+1}^{n} (t_1 - Qt_j) \\ &= (-Q^{-1})^{(s-1)} \prod_{1\leqq i<j\leqq n} (t_{\sigma_s(i)} - Qt_{\sigma_s(j)}). \end{aligned} \tag{4.24}$$

The substitution

$$t_i \to t_{\pi_s(i)}, \qquad 1\leqq i\leqq n, \tag{4.25}$$

gives

$$\begin{aligned} &\int \cdots \int f(\mathbf{t}) \prod_{j=2}^{s} \frac{(t_1 - Q^{-1}t_j)}{(t_1 - Qt_j)} \prod_{1\leqq i<j\leqq n} (t_i - Qt_j)\, d_\mu(\mathbf{t}) \\ &\quad = Q^{-(s-1)} \int \cdots \int f(\pi_s(\mathbf{t})) \prod_{1\leqq i<j\leqq n} (t_i - Qt_j)\, d_\mu(\mathbf{t}). \end{aligned} \tag{4.26}$$

We again assume that (2.6) holds. Setting $s=v-1$, $Q=q^k$, in (4.26), our results (3.7), (3.8), and (3.9) become

$$\begin{aligned} {}_q^1A_{n,m}^v(x, y, k) &= \frac{q^{-k(v-2)}}{(1+q^k)} \int_0^{1/q} \cdots \int_0^{1/q} \prod_{i=1}^{m} t_i\, {}_qW_n(x, y, k)\, d_qt_1 \cdots d_qt_n \\ &= \frac{q^{-k(v-2)}}{(1+q^k)}\, {}_qS_{n,m}(x, y, k), \qquad 2\leqq v\leqq m, \end{aligned} \tag{4.27}$$

$$
\begin{aligned}
{}_q^0A_{n,m}^v(x,y,k) &= \frac{q^{-k(v-2)}}{(1+q^k)}\int_0^{1/q}\cdots\int_0^{1/q}\prod_{i=1}^{m-1} t_i\,{}_qW_n(x,y,k)\,d_qt_1\cdots d_qt_n \\
&= \frac{q^{-k(v-2)}}{(1+q^k)}\,{}_qS_{n,m-1}(x,y,k), \qquad m<v\leqq n,
\end{aligned} \tag{4.28}
$$

$$
\begin{aligned}
{}_q^1A_{n,m}^v(x,y,k) &= \frac{q^{-k(v-2)}}{(1+q^{2k})}\int_0^{1/q}\cdots\int_0^{1/q}\prod_{i=1}^{m-1} t_i(t_{v-1}+q^kt_v) \\
&\quad\cdot\,{}_qW_n(x,y,k)\,d_qt_1\cdots d_qt_n, \qquad m<v\leqq n.
\end{aligned} \tag{4.29}
$$

We have

$$
\begin{aligned}
e([1,m-1]\cup\{v-1\}) &= v-m-1, \\
e([1,m-1]\cup\{v\}) &= v-m.
\end{aligned} \tag{4.30}
$$

Setting $Q=q^k$ in (4.14) and using (4.30), (4.29) becomes

$$
\begin{aligned}
{}_q^1A_{n,m}^v(x,y,k) &= \frac{q^{-k(v-2)}}{(1+q^{2k})}(q^{k(v-m-1)}+q^{k(v-m+1)})\,{}_qS_{n,m}(x,y,k) \\
&= q^{k(1-m)}\,{}_qS_{n,m}(x,y,k), \qquad m<v\leqq n.
\end{aligned} \tag{4.31}
$$

Set $s=n$, $Q=q^k$, in (4.26). Recalling (2.16), we obtain

$$
{}_q^\delta K_{n,m}(x,y,k) = q^{-k(n-1)}\int_0^{1/q}\cdots\int_0^{1/q} t_n^\delta\prod_{i=1}^{m-1} t_i\,{}_qW_n(x,y,k)\,d_qt_1\cdots d_qt_n. \tag{4.32}
$$

We have

$$
{}_q^0K_{n,m}(x,y,k) = q^{-k(n-1)}\,{}_qS_{n,m-1}(x,y,k). \tag{4.33}
$$

Since

$$
e([1,m-1]\cup\{n\}) = n-m, \tag{4.34}
$$

(4.14) gives

$$
{}_q^1K_{n,m}(x,y,k) = q^{k(1-m)}\,{}_qS_{n,m}(x,y,k). \tag{4.35}
$$

5. A proof of Theorem 2 for $l=0$. We must first give a q-analogue of (1.10) Assume that (2.6) holds. Set $\delta=0$ in (2.14) and use (3.6), (4.28) and (4.33). We obtair

$$
\begin{aligned}
0 &= \frac{(1-q^{2k})}{(1-q)}\sum_{v=m+1}^{n}\frac{q^{k(v-2)}}{(1+q^k)}\,{}_qS_{n,m-1}(x,y,k) \\
&\quad + q^{k(n-1)}\frac{(1-q^x)}{(1-q)}\,{}_qS_{n,m-1}(x,y,k) \\
&\quad - q^{2k(n-1)+x+1}\frac{(1-q^{(y-1)})}{(1-q)}\,{}_q^0E_{n,m}(x,y,k).
\end{aligned} \tag{5.1}
$$

Multiply by $(1-q)$ and move the term with ${}_q^0E_{n,m}(x,y,k)$ to the left side. This gives

$$
\begin{aligned}
&q^{2k(n-1)+x+1}(1-q^{(y-1)})\,{}_q^0E_{n,m}(x,y,k) \\
&= \left[(1-q^k)\sum_{v=m+1}^{n} q^{k(v-2)}+q^{k(n-1)}(1-q^x)\right]{}_qS_{n,m-1}(x,y,k) \\
&= [(q^{k(m-1)}-q^{k(n-1)})+q^{k(n-1)}(1-q^x)]\,{}_qS_{n,m-1}(x,y,k) \\
&= q^{k(m-1)}(1-q^{x+k(n-m)})\,{}_qS_{n,m-1}(x,y,k).
\end{aligned} \tag{5.2}
$$

Set $\delta = 0$ in (3.10), solve for $q\,{}_q^1E_{n,m}(x, y, k)$, and use (4.35). This gives

$$\begin{aligned} q\,{}_q^1E_{n,m}(x, y, k) &= -{}_q^1K_{n,m}(x, y, k) + {}_q^0E_{n,m}(x, y, k) \\ &= -q^{k(1-m)}\,{}_qS_{n,m}(x, y, k) + {}_q^0E_{n,m}(x, y, k). \end{aligned} \tag{5.3}$$

Set $\delta = 1$ in (2.14) and use (4.27), (4.31), (4.35), and (5.3). We obtain

$$\begin{aligned} 0 = {} & \frac{(1-q^{2k})}{(1-q)} \sum_{v=2}^{m} \frac{q^{k(v-2)}}{(1+q^k)}\,{}_qS_{n,m}(x, y, k) \\ & + \frac{(1-q^{2k})}{(1-q)} \sum_{v=m+1}^{n} q^{k(2v-m-3)}\,{}_qS_{n,m}(x, y, k) \\ & + q^{k(2n-m-1)} \frac{(1-q^{x+1})}{(1-q)}\,{}_qS_{n,m}(x, y, k) \\ & + q^{k(2n-m-1)+x+1} \frac{(1-q^{(y-1)})}{(1-q)}\,{}_qS_{n,m}(x, y, k) \\ & - q^{2k(n-1)+x+1} \frac{(1-q^{(y-1)})}{(1-q)}\,{}_q^0E_{n,m}(x, y, k). \end{aligned} \tag{5.4}$$

Multiply by $(1-q)$ and move the term with ${}_q^0E_{n,m}(x, y, k)$ to the left side. This yields

$$\begin{aligned} & q^{2k(n-1)+x+1}(1-q^{(y-1)})\,{}_q^0E_{n,m}(x, y, k) \\ & = \Bigg[(1-q^k) \sum_{v=2}^{m} q^{k(v-2)} + (1-q^{2k}) \sum_{v=m+1}^{n} q^{k(2v-m-3)} \\ & \quad + q^{k(2n-m-1)}(1-q^{x+1}) + q^{k(2n-m-1)+x+1}(1-q^{(y-1)}) \Bigg]\,{}_qS_{n,m}(x, y, k) \\ & = [(1-q^{k(m-1)}) + (q^{k(m-1)} - q^{k(2n-m-1)}) \\ & \quad + (q^{k(2n-m-1)} - q^{k(2n-m-1)+x+y})]\,{}_qS_{n,m}(x, y, k) \\ & = (1-q^{x+y+k(2n-m-1)})\,{}_qS_{n,m}(x, y, k). \end{aligned} \tag{5.5}$$

Comparing (5.2) and (5.5), we have

$$q^{k(m-1)}(1-q^{x+k(n-m)})\,{}_qS_{n,m-1}(x, y, k) = (1-q^{x+y+k(2n-m-1)})\,{}_qS_{n,m}(x, y, k) \tag{5.6}$$

or

$${}_qS_{n,m}(x, y, k) = q^{k(m-1)} \frac{(1-q^{x+k(n-m)})}{(1-q^{x+y+k(2n-m-1)})}\,{}_qS_{n,m-1}(x, y, k), \tag{5.7}$$

provided (2.6) holds. It is easy to see that

$${}_qpr_{n,m}(x, y, k) = q^{[kx\binom{n}{2}+k\binom{m}{2}]} \prod_{i=1}^{n} \frac{\Gamma_q(x+(n-i)k+\chi(i \leqq m))}{\Gamma_q(x+y+(2n-i-1)k+\chi(i \leqq m))} \tag{5.8}$$

satisfies (5.7) since, by (1.7),

$$\Gamma_q(x+1) = \frac{(1-q^x)}{(1-q)}\Gamma_q(x). \tag{5.9}$$

Clearly

$${}_qS_{n,n}(x, y, k) = {}_qS_{n,0}(x+1, y, k) \tag{5.10}$$

and ${}_qpr_{n,m}(x, y, k)$ also satisfies (5.10). Hence

$$ {}_qQ_{n,m}(x, y, k) = \frac{{}_qS_{n,m}(x, y, k)}{{}_qpr_{n,m}(x, y, k)} \tag{5.11} $$

satisfies

$$ \begin{aligned} {}_qQ_{n,m}(x, y, k) &= {}_qQ_{n,m-1}(x, y, k), \\ {}_qQ_{n,n}(x, y, k) &= {}_qQ_{n,0}(x+1, y, k). \end{aligned} \tag{5.12} $$

We easily obtain

$$ {}_qQ_{n,m}(x, y, k) = {}_qQ_{n,m}(x+1, y, k) \tag{5.13} $$

for all m, $0 \leqq m \leqq n$. Thus ${}_qQ_{n,m}(x, y, k)$ is independent of m and periodic in x with period 1. We extend ${}_qQ_{n,m}(x, y, k)$ to all x.

In order to prove Theorem 2 for $l = 0$, we must show that

$$ {}_qQ_{n,m}(x, y, k) = q^{2k^2\binom{n}{3}} \prod_{i=1}^{n} \Gamma_q(y+(n-i)k) \frac{\Gamma_q(1+ik)}{\Gamma_q(1+k)}. \tag{5.14} $$

By (5.12), it suffices to treat the case $m = 0$. We proceed by induction on n. For $n = 1$, (1.4) is the q-analogue

$$ \int_0^1 t^{(x-1)} \frac{(qt)_\infty}{(q^y t)_\infty} \, d_q t = \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)} \tag{5.15} $$

of the beta integral. See Askey [5].

It is clear from (5.9) that

$$ \lim_{x \to 0} \frac{(1-q^x)}{(1-q)} \Gamma_q(x) = \Gamma_q(1) = 1 \tag{5.16} $$

and from (5.11) that

$$ {}_qS_{n,m}(x, y, k) = {}_qpr_{n,m}(x, y, k) \; {}_qQ_{n,m}(x, y, k). \tag{5.17} $$

Observe that ${}_qpr_{n,0}(x, y, k)$ has $\Gamma_q(x)$ as a factor. A simple computation gives

$$ \begin{aligned} \lim_{x \to 0} \frac{(1-q^x)}{(1-q)} {}_qS_{n,0}(x, y, k) &= \lim_{x \to 0} \frac{(1-q^x)}{(1-q)} {}_qpr_{n,0}(x, y, k) \; {}_qQ_{n,0}(x, y, k) \\ &= \frac{1}{\Gamma_q(y+(n-1)k)} \prod_{i=1}^{n-1} \frac{\Gamma_q((n-i)k)}{\Gamma_q(y+(2n-i-1)k)} \, {}_qQ_{n,0}(0, y, k). \end{aligned} \tag{5.18} $$

However, the recurrence relation (4.21) and our induction hypothesis give

$$ \begin{aligned} \lim_{x \to 0} \frac{(1-q^x)}{(1-q)} {}_qS_{n,0}(x, y, k) &= \frac{(1-q^{kn})}{(1-q^k)} {}_qS_{n-1,0}(2k, y, k) \\ &= \frac{(1-q^{kn})}{(1-q^k)} q^{[2k^2\binom{n-1}{2}+2k^2\binom{n-1}{3}]} \\ &\quad \cdot \prod_{i=1}^{n-1} \frac{\Gamma_q(2k+(n-1-i)k)\Gamma_q(y+(n-1-i)k)\Gamma_q(1+ik)}{\Gamma_q(2k+y+(2(n-1)-i-1)k)\Gamma_q(1+k)}. \end{aligned} \tag{5.19} $$

Equating these two results and solving for ${}_qQ_{n,0}(0, y, k)$, we obtain

$$
\begin{aligned}
{}_qQ_{n,0}(0,y,k) &= \frac{(1-q^{kn})}{(1-q^k)} q^{[2k^2\binom{n-1}{2}+2k^2\binom{n-1}{3}]} \Gamma_q(y+(n-1)k) \prod_{i=1}^{n-1} \Gamma_q(y+(n-1-i)k) \\
&\quad \cdot \prod_{i=1}^{n-1} \frac{\Gamma_q((n+1-i)k)}{\Gamma_q((n-i)k)} \frac{\Gamma_q(1+ik)}{\Gamma_q(1+k)} \\
&= q^{2k^2\binom{n}{3}} \prod_{i=1}^{n} \Gamma_q(y+(n-i)k) \frac{(1-q^{kn})}{(1-q^k)} \frac{\Gamma_q(nk)}{\Gamma_q(k)} \prod_{i=1}^{n-1} \frac{\Gamma_q(1+ik)}{\Gamma_q(1+k)} \\
&= q^{2k^2\binom{n}{3}} \prod_{i=1}^{n} \Gamma_q(y+(n-i)k) \frac{\Gamma_q(1+ik)}{\Gamma_q(1+k)}.
\end{aligned} \tag{5.20}
$$

By (5.13), this establishes (5.14) when x is an integer. To treat all x, we may use Liouville's theorem: a bounded entire function is constant. See Ahlfors [1, Chap. 4]. Let $a, b \in R$. Since $0<q<1$, we have

$$
|q^{a+ib}| = q^a, \qquad |(1-q)^{a+ib}| = (1-q)^a, \tag{5.21}
$$

and

$$
(q^a)_\infty \leqq |(q^{a+ib})_\infty| \leqq (-q^a)_\infty, \qquad a>0. \tag{5.22}
$$

This gives

$$
|{}_qS_{n,0}(x,y,k)| \leqq {}_qS_{n,0}(\mathrm{Re}\,(x), \mathrm{Re}\,(y), k). \tag{5.23}
$$

Recalling (1.7), we have

$$
(1-q)^{(1-a)} \frac{(q)_\infty}{(-q^a)_\infty} \leqq |\Gamma_q(a+ib)| \leqq (1-q)^{(1-a)} \frac{(q)_\infty}{(q^a)_\infty} = \Gamma_q(a), \qquad a>0. \tag{5.24}
$$

This shows that the modulus of each factor of ${}_qpr_{n,0}(x, y, k)$ (5.8) is bounded above and below by a continuous function of $\mathrm{Re}\,(x)$. Using (5.11) and (5.23), we see that $|{}_qQ_{n,0}(x, y, k)|$ is bounded by a continuous function of $\mathrm{Re}\,(x)$. Since a continuous function assumes its maximum on a closed set, $|{}_qQ_{n,0}(x, y, k)|$ is bounded for all x in the strip

$$
S = \{x \mid 1 \leqq \mathrm{Re}\,(x) \leqq 2\}. \tag{5.25}
$$

By (5.13), it is bounded for all x. The result (5.14) now follows by Liouville's theorem. The restriction $\mathrm{Re}\,(y)>1$ (2.6) is easily removed by analytic continuation. This completes the proof of the case $l=0$ of Theorem 2.

6. A proof of Theorem 2. By (4.19), the case $l=0$ of Theorem 2 (1.4) gives the case $l=0$ of Conjectures 6 and 11 of [10]. These conjectures now follow by Theorems 10 and 12 of [10]. A similar analysis establishes Theorem 2.

We have

$$
\begin{aligned}
{}_qS_{n,m,l}(x,y,k) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{m} t_i \prod_{i=n-l+1}^{n} (1-q^y t_i)\, {}_qW_n(x,y,k)\, d_qt_1 \cdots d_q t_n \\
&= \sum_{j=0}^{l} (-q^y)^j \sum_{\substack{A\subset[n-l+1,n] \\ |A|=j}} \\
&\quad \cdot \int_0^1 \cdots \int_0^1 \prod_{i=1}^{m} t_i \prod_{i\in A} t_i\, {}_qW_n(x,y,k)\, d_qt_1 \cdots d_qt_n.
\end{aligned} \tag{6.1}
$$

Observe that

$$e([1, m] \cup A) = \sum_{i \in A} i - (m+1)|A| - \binom{|A|}{2}. \tag{6.2}$$

Setting $Q = q^k$ in (4.14) and using (6.2), (6.1) becomes

$$ {}_qS_{n,m,l}(x, y, k) = \sum_{j=0}^{l} (-q^y)^j \sum_{\substack{A \subset [n-l+1,n] \\ |A|=j}} q^{[k(\sum_{i \in A} i-(m+1)j-\binom{j}{2})]}\, {}_qS_{n,m+j}(x, y, k). \tag{6.3}$$

We can perform the inner sum by using the well-known (see Andrews [2, (2.2.1)]) q-binomial theorem

$$\frac{(at)_\infty}{(t)_\infty} = \sum_{j=0}^{\infty} \frac{(a)_j}{(q)_j} t^j, \qquad |t| < 1. \tag{6.4}$$

For $a = q^{-l}$, $t = xq^{n+1}$, this gives

$$(xq^{n-l+1})_l = \sum_{j=0}^{l} \frac{(q^{-l})_j}{(q)_j} (xq^{n+1})^j. \tag{6.5}$$

Equating coefficients yields

$$\sum_{\substack{A \subset [n-l+1,n] \\ |A|=j}} q^{\sum_{i \in A} i} = (-1)^j \frac{(q^{-l})_j}{(q)_j} q^{(n+1)j}. \tag{6.6}$$

Replacing q by q^k in (6.6) and substituting into (6.3), we obtain

$$ {}_qS_{n,m,l}(x, y, k) = \sum_{j=0}^{l} q^{(yj+k(n-m)j-k\binom{j}{2})} \frac{(q^{-lk}; q^k)_j}{(q^k; q^k)_j}\, {}_qS_{n,m+j}(x, y, k). \tag{6.7}$$

Since we have evaluated each of the integrals on the right side of (6.7), we can evaluate ${}_qS_{n,m,l}(x, y, k)$. By (5.9) and (1.4), we have

$$\begin{aligned} \frac{{}_qS_{n,m+j}(x, y, k)}{{}_qS_{n,m}(x, y, k)} &= q^{(kmj+k\binom{j}{2})} \prod_{i=m+1}^{m+j} \frac{(1-q^{x+(n-i)k})}{(1-q^{x+y+(2n-i-1)k})} \\ &= q^{(-yj+k(m-n+1)j+k\binom{j}{2})} \frac{(q^{-x-(n-m-1)k}; q^k)_j}{(q^{-x-y-(2n-m-2)k}; q^k)_j}. \end{aligned} \tag{6.8}$$

Equation (6.7) now gives

$$ {}_qS_{n,m,l}(x, y, k) = {}_qS_{n,m}(x, y, k) \sum_{j=0}^{l} \frac{(q^{-lk}; q^k)_j (q^{-x-(n-m-1)k}; q^k)_j}{(q^k; q^k)_j (q^{-x-y-(2n-m-2)k}; q^k)_j} q^{kj}. \tag{6.9}$$

This is the same series that arose [10, (8.13)] in our analysis of ${}_qI_{n,m,l}(x, y, k)$. It can be summed by the special case (see Slater [16, (3.3.2.7)])

$$b^n \frac{(d/b)_n}{(d)_n} = \sum_{i=0}^{n} \frac{(q^{-n})_i (b)_i}{(q)_i (d)_i} q^i \tag{6.10}$$

of the well-known q-analogue [16, (3.3.2.2)] of Saalschütz's theorem. Taking base q^k in (6.10), we obtain

$$ {}_qS_{n,m,l}(x, y, k) = {}_qS_{n,m}(x, y, k) \frac{(q^{y+(n-l)k}; q^k)_l}{(q^{x+y+(2n-m-l-1)k}; q^k)_l}. \tag{6.11}$$

Theorem 2 (1.4) now follows easily using (5.9).

7. A proof of Theorem 3. Since k is a nonnegative integer, we have the finite sum

$$\prod_{1\leqq i<j\leqq n} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k} = \sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} t_i^{\alpha_i}, \tag{7.1}$$

where $\alpha_i \geqq 0$, $1 \leqq i \leqq n$,

$$\sum_{i=1}^{n} \alpha_i = 2k\binom{n}{2}, \tag{7.2}$$

and each ${}_qc_n^k(\alpha)$ is a polynomial in q. Substitute (7.1) into (1.4) and use (5.15) and (5.9). This yields

$$\begin{aligned} {}_qS_{n,m,l}(x, y, k) &= \int_0^1 \cdots \int_0^1 \prod_{i=1}^{n} t_i^{(x-1)+\chi(i\leqq m)} \frac{(qt_i)_\infty}{(q^{y+\chi(n-i+1\leqq l)}t_i)_\infty} \\ &\quad \cdot \left(\sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} t_i^{\alpha_i}\right) d_qt_1 \cdots d_qt_n \\ &= \sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} \frac{\Gamma_q(x+\chi(i\leqq m)+\alpha_i)\Gamma_q(y+\chi(n-i+1\leqq l))}{\Gamma_q(x+y+\chi(i\leqq m)+\chi(n-i+1\leqq l)+\alpha_i)} \\ &= \prod_{i=1}^{n} \frac{\Gamma_q(x+\chi(i\leqq m))\Gamma_q(y+\chi(n-i+1\leqq l))}{\Gamma_q(x+y+\chi(i\leqq m)+\chi(n-i+1\leqq l))} \\ &\quad \cdot \sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} \frac{(q^{x+\chi(i\leqq m)})_{\alpha_i}}{(q^{x+y+\chi(i\leqq m)+\chi(n-i+1\leqq l)})_{\alpha_i}}. \end{aligned} \tag{7.3}$$

Equate (7.3) and Theorem 2 (1.4) and solve for the last $\sum_\alpha$. Using (5.9), we obtain

$$\begin{aligned} \sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} \frac{(q^{x+\chi(i\leqq m)})_{\alpha_i}}{(q^{x+y+\chi(i\leqq m)+\chi(n-i+1\leqq l)})_{\alpha_i}} &= q^{[kx\binom{n}{2}+k\binom{m}{2}+2k^2\binom{n}{3}]} \\ \cdot \prod_{i=1}^{n} \frac{(q^{x+\chi(i\leqq m)})_{(n-i)k}(q^{y+\chi(i\leqq l)})_{(n-i)k}(q)_{ik}}{(q^{x+y+\chi(i\leqq m+l)})_{(2n-i-1)k}(q)_k}. \end{aligned} \tag{7.4}$$

Since the left side of (7.4) is a finite sum, both sides are rational functions of q, q^x, and q^y. Thus (7.4) holds for all x and y. Observe that we do not need to use Liouville's theorem or analytic continuation to prove Theorem 2.

By reversing the order of the factors, we have

$$(x)_n = (-x)^n q^{\binom{n}{2}}\left(\frac{q^{1-n}}{x}\right)_n. \tag{7.5}$$

Applying this to $(t_i/t_j)_k$ gives

$$\left(q\frac{t_j}{t_i}\right)_k\left(\frac{t_i}{t_j}\right)_k = \left(-\frac{1}{t_it_j}\right)^k q^{\binom{k}{2}} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k}. \tag{7.6}$$

Multiplying these equations for $1 \leqq i < j \leqq n$ and using (7.1), we obtain

$$\begin{aligned} \prod_{1\leqq i<j\leqq n}\left(q\frac{t_j}{t_i}\right)_k\left(\frac{t_i}{t_j}\right)_k &= (-1)^{k\binom{n}{2}} q^{\binom{k}{2}\binom{n}{2}} \prod_{i=1}^{n} t_i^{-(n-1)k} \prod_{1\leqq i<j\leqq n} t_i^{2k}\left(q^{1-k}\frac{t_j}{t_i}\right)_{2k} \\ &= (-1)^{k\binom{n}{2}} q^{\binom{k}{2}\binom{n}{2}} \sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} t_i^{(\alpha_i-(n-1)k)}. \end{aligned} \tag{7.7}$$

Recall (Kadell [9, (3.31)]) that

$$(qt)_a\left(\frac{1}{t}\right)_b = \sum_{j=-b}^{a} (-t)^j q^{\binom{j+1}{2}} \frac{(q)_{a+b}}{(q)_{a-j}(q)_{b+j}} \tag{7.8}$$

is equivalent to the special case (6.5) of the q-binomial theorem. Replacing t by t/q gives

$$(t)_a\left(\frac{q}{t}\right)_b = \sum_{j=-b}^{a} (-t)^j q^{\binom{j}{2}} \frac{(q)_{a+b}}{(q)_{a-j}(q)_{b+j}}. \tag{7.9}$$

Substitute (7.7) into (1.12) and extract the constant term using (7.8) and (7.9). This yields

$$\begin{aligned} {}_qCS_{n,m,l}(a,b,k) &= (-1)^{k\binom{n}{2}} q^{\binom{k}{2}\binom{n}{2}} \\ &\cdot \sum_{\boldsymbol{\alpha}} {}_qc_n^k(\boldsymbol{\alpha}) \prod_{i=1}^{n} (-1)^{(\alpha_i-(n-1)k)} q^{(\chi(i>m)-\frac{\alpha_i}{2}+(n-1)k)} \\ &\cdot \prod_{i=1}^{n} \frac{(q)_{a+b+\chi(n-i+1\leqq l)}}{(q)_{a+\chi(i\leqq m)+\chi(n-i+1\leqq l)-(n-1)k+\alpha_i}(q)_{b-\chi(i\leqq m)+(n-1)k-\alpha_i}}. \end{aligned} \tag{7.10}$$

By (7.2), we have

$$\prod_{i=1}^{n} (-1)^{(\alpha_i-(n-1)k)} = \prod_{i=1}^{n} (-1)^{\alpha_i} = 1. \tag{7.11}$$

Equation (7.5) gives

$$\begin{aligned} \frac{1}{(q)_{b-\chi(i\leqq m)+(n-1)k-\alpha_i}} &= \frac{(q^{b+1-\chi(i\leqq m)+(n-1)k-\alpha_i})_{\alpha_i}}{(q)_{b-\chi(i\leqq m)+(n-1)k}} \\ &= (-1)^{\alpha_i} q^{\alpha_i(b+1-\chi(i\leqq m)+(n-1)k-\alpha_i)+\binom{\alpha_i}{2}} \\ &\cdot \frac{(q^{-b+\chi(i\leqq m)-(n-1)k})_{\alpha_i}}{(q)_{b-\chi(i\leqq m)+(n-1)k}}. \end{aligned} \tag{7.12}$$

Using

$$\binom{A}{2} = -A + \binom{1+A}{2}, \tag{7.13}$$

$$\binom{A+B}{2} = \binom{A}{2} + AB + \binom{B}{2}, \tag{7.14}$$

$$\binom{1-A}{2} = \binom{A}{2}, \tag{7.15}$$

and (7.2), we obtain

$$\prod_{i=1}^{n} q^{(\chi(i>m)-\frac{\alpha_i}{2}+(n-1)k)+\alpha_i(b+1-\chi(i\leqq m)+(n-1)k-\alpha_i)+\binom{\alpha_i}{2}} = q^{(n-m)(n-1)k+2bk\binom{n}{2}+n\binom{(n-1)k}{2}} \tag{7.16}$$

Equation (7.10) now becomes

$$\begin{aligned} {}_qCS_{n,m,l}(a,b,k) &= (-1)^{k\binom{n}{2}} q^{\binom{k}{2}\binom{n}{2}+(n-m)(n-1)k+2bk\binom{n}{2}+n\binom{(n-1)k}{2}} \\ &\cdot \prod_{i=1}^{n} \frac{(q)_{a+b+\chi(i\leqq l)}}{(q)_{a+\chi(i\leqq m+l)-(n-1)k}(q)_{b-\chi(i\leqq m)+(n-1)k}} \\ &\cdot \sum_{\boldsymbol{\alpha}} {}_qc_n^k(\boldsymbol{\alpha}) \prod_{i=1}^{n} \frac{(q^{-b+\chi(i\leqq m)-(n-1)k})_{\alpha_i}}{(q^{a+1+\chi(i\leqq m)+\chi(n-i+1\leqq l)-(n-1)k})_{\alpha_i}}. \end{aligned} \tag{7.17}$$

We can evaluate the last $\sum_{\alpha}$ by taking

$$x=-b-(n-1)k, \qquad y=a+b+1, \tag{7.18}$$

in (7.4). This gives

$$\begin{aligned}\sum_{\alpha} {}_qc_n^k(\alpha) \prod_{i=1}^{n} \frac{(q^{-b+\chi(i\leqq m)-(n-1)k})_{\alpha_i}}{(q^{a+1+\chi(i\leqq m)+\chi(n-i+1\leqq l)-(n-1)k})_{\alpha_i}} \\ = q^{-k(b+(n-1)k)\binom{n}{2}+k\binom{m}{2}+2k^2\binom{n}{3}} \\ \cdot \prod_{i=1}^{n} \frac{(q^{-b+\chi(i\leqq m)-(n-1)k})_{(n-i)k}(q^{a+b+1+\chi(i\leqq l)})_{(n-i)k}(q)_{ik}}{(q^{a+1-(n-1)k+\chi(i\leqq m+l)})_{(2n-i-1)k}(q)_k}.\end{aligned} \tag{7.19}$$

By (7.5), we have

$$\begin{aligned}(q^{-b+\chi(i\leqq m)-(n-1)k})_{(n-i)k} = (-1)^{(n-i)k} q^{(-b+\chi(i\leqq m)-(n-1)k)(n-i)k+\binom{(n-i)k}{2}} \\ \cdot (q^{b+1-\chi(i\leqq m)+(i-1)k})_{(n-i)k}.\end{aligned} \tag{7.20}$$

Clearly, for fixed j,

$$\sum_{i=1}^{n} \binom{n-i}{j-1} = \binom{n}{j}. \tag{7.21}$$

Substituting (7.19) into (7.17) and simplifying by (7.20) and (7.21), we obtain

$${}_qCS_{n,m,l}(a,b,k) = q^{\mathrm{EXP}} \prod_{i=1}^{n} \frac{(q)_{a+b+(n-i)k+\chi(i\leqq l)}(q)_{ik}}{(q)_{a+(n-i)k+\chi(i\leqq m+l)}(q)_{b+(i-1)k-\chi(i\leqq m)}(q)_k}, \tag{7.22}$$

where

$$\begin{aligned}\mathrm{EXP} = \binom{k}{2}\binom{n}{2} + (n-m)(n-1)k + n\binom{(n-1)k}{2} - 2k^2(n-1)\binom{n}{2} \\ + k\binom{m}{2} + 2k^2\binom{n}{3} + \sum_{i=1}^{m}(n-i)k + \sum_{i=1}^{n}\binom{(n-i)k}{2}.\end{aligned} \tag{7.23}$$

It is easy to see that EXP is independent of m. Thus

$$\begin{aligned}\mathrm{EXP} = \binom{k}{2}\binom{n}{2} + 2k\binom{n}{2} + n\binom{(n-1)k}{2} - 2k^2(n-1)\binom{n}{2} + 2k^2\binom{n}{3} \\ + \sum_{i=1}^{n}\binom{(n-i)k}{2}.\end{aligned} \tag{7.24}$$

Observe that

$$\binom{AB}{2} = A^2\binom{B}{2} + \binom{A}{2}B. \tag{7.25}$$

Take $A=k$, $B=n-i$, in (7.25) and $j=2,3$, in (7.21). This gives

$$\sum_{i=1}^{n}\binom{(n-i)k}{2} = \sum_{i=1}^{n} k^2\binom{n-i}{2} + \binom{k}{2}(n-i) = k^2\binom{n}{3} + \binom{k}{2}\binom{n}{2}. \tag{7.26}$$

We obtain

$$\text{EXP} = 2\binom{k}{2}\binom{n}{2} + 2k\binom{n}{2} + n\left(k^2\binom{n-1}{2} + \binom{k}{2}(n-1)\right) - 2k^2(n-1)\binom{n}{2} + 3k^2\binom{n}{3}$$

$$= k\binom{n}{2}[(k-1)+2+k(n-2)+(k-1)-2k(n-1)+k(n-2)] \tag{7.27}$$

$$= 0.$$

This proves Theorem 3 since (7.22) now agrees with (1.14).

A similar argument shows that Theorem 2 follows from Theorem 3. Indeed, Theorems 2 and 3 are equivalent to the summation formula (7.4). Observe that x in (7.18) is negative. Thus, (7.4) provides a common analytic continuation of Theorem 2 (1.4) and Theorem 3 (1.14).

REFERENCES

[1] L. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, 1966.

[2] G. E. ANDREWS, *The Theory of Partitions*, Addison-Wesley, Reading, MA, 1976.

[3] ———, *q-Series*, Regional Conference in Pure Mathematics, American Mathematical Society, Providence, RI, 1986.

[4] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.

[5] R. ASKEY, *The q-gamma and q-beta functions*, Applicable Anal., 8 (1978), pp. 125-141.

[6] ———, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.

[7] I. P. GOULDEN AND D. M. JACKSON, *Combinatorial Enumeration*, John Wiley, New York, 1983.

[8] F. H. JACKSON, *On q-definite integrals*, Quart. J. Pure Appl. Math., 41 (1910), pp. 193-203.

[9] K. W. J. KADELL, *A proof of Andrews' q-Dyson conjecture for $n = 4$*, Trans. Amer. Math. Soc., 290 (1985), pp. 127-144.

[10] ———, *A proof of some q-analogues of Selberg's integral for $k = 1$*, SIAM J. Math. Anal., 19 (1988), pp. 944-968.

[11] M. G. KENDALL AND A. STUART, *The Advanced Theory of Statistics*, 2, Hafner, New York, 1973.

[12] I. G. MACDONALD, *Some conjectures for root systems and finite reflection groups*, SIAM J. Math. Anal., 13 (1982), pp. 988-1007.

[13] P. A. MACMAHON, *Two applications of general theorems in combinatory analysis*, Proc. London Math. Soc. (2), 15 (1916), pp. 314-321.

[14] W. MORRIS, *Constant term identities for finite and affine root systems*, Ph.D. thesis, Univ. of Wisconsin, Madison, WI, 1982.

[15] A. SELBERG, *Bemerkninger om et multipelt integral*, Nordisk Tidskr., 26 (1944), pp. 71-78.

[16] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, London, 1966.

SIAM J. MATH. ANAL.
Vol. 19, No. 4, July 1988

A UNIFIED APPROACH TO MACDONALD'S ROOT-SYSTEM CONJECTURES*

DORON ZEILBERGER†

Dedicated to Dennis Stanton and John Stembridge for reminding me that antisymmetry is even more powerful than symmetry.

"*Yes, of course. It works with herring, but will it work with ferrous metals?*"
(*Woody Allen* [Al]).

Abstract. Using ideas of Stembridge and Stanton a method is presented that should settle the Macdonald (and the more refined Macdonald-Morris) root-system conjectures for any *specific* root system, provided there is sufficient computer time, memory space, and (for now) some luck. The method consists of an algorithm that reduces Macdonald's conjecture for a given root system to a finite, albeit long, algebraic calculation, which is then performed using computer algebra. The method is illustrated by proving the so far open G_2^{ν} case of the Macdonald-Morris conjectures. The question that remains is: will it work with E_8 (and F_4, E_6, E_7)?

Key words. Macdonald's root-system conjectures, constant term, q-analogue, computer algebra

AMS(MOS) subject classifications. 05A15, 05A17, 33A15, 33A75

Introduction. This paper is about Macdonald's *root-system* conjectures. In order to understand it, it is necessary to know a little bit about root systems and their Weyl groups. While it seems obvious that before one can talk about *root-system* conjectures one has to know about root systems, this is not the case for many of the papers on this subject. By the classification theorem for root systems, it is possible to spell out what the conjectures say for each of the four infinite families and the five exceptional root systems, and then treat each case separately [Mo]. Although only one root-system is treated at a time in this paper, its method is cast in the general root-system mold.

Historically root systems first came up in the deep and sophisticated theory of Lie algebras. This noble birth gave them a fancy aura that scared away many a plebeian mathematician. However, root systems are really very simple-minded, combinatorial-geometrical structures and it is possible, perhaps even preferable, to study root systems without knowing anything about Lie algebras.

A root system is a finite collection of vectors, called roots, in regular (Euclidean) space such that if you place a mirror perpendicular to any of them, the image of the visible part that is reflected in the mirror coincides exactly with the invisible part behind the mirror. Furthermore, the vector difference between any root and its image under any such mirror is an *integer* multiple of the root corresponding to the mirror (i.e., the root that is perpendicular to the mirror). These two conditions are very strong and it turns out (the classification theorem) that all irreducible root systems fall into five infinite families and five exceptionals. If you add the condition that these vectors can only be parallel to their negatives (reduced root systems) then one infinite family (BC_n) drops out.

* Received by the editors March 2, 1987; accepted for publication August 4, 1987. This research was partially supported by the National Science Foundation.

† Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

An excellent treatment of root systems and Weyl groups is given in Chapters 2 and 10 of Carter's book [C]. These two chapters are completely independent of the rest of the book and are entirely elementary. This paper can be understood by any one who has read the first two sections of Chapter 2 and the first two sections of Chapter 10 of [C]. A comprehensive and (surprisingly) quite readable account is given in [Bo], but for the present paper [C] is more than enough.

Notation. The Macdonald conjectures are about certain multivariable Laurent polynomials. A Laurent polynomial is a linear combination of monomials that may have negative integer exponents as well as positive integer exponents. For example $x+1+x^{-1}$ is a Laurent polynomial in one variable and $x+y+x^{-1}y^2$ is one in two variables. Usually x denotes a vector of variables, $x=(x_1,\cdots,x_l)$ and α a vector of integers, $\alpha=(\alpha_1,\cdots,\alpha_l)$. Also

$$x^{\alpha}=x_1^{\alpha_1}\cdots x_l^{\alpha_l}.$$

For example $x^{(1,-2,5)}=x_1x_2^{-2}x_3^5$.

For the roots α of a root system, x^{α}, are often called "formal exponentials." But since all root systems can be made to have all their roots with integer components, these exponentials can be easily defrocked of their formality. The root lattice of a root system consists of all integer linear combination of roots, and all our Laurent polynomials will be linear combinations of monomials x^{γ} for γ in the root lattice. The Weyl group W of a root system [C, Chap. 2] acts on the roots, and by linearity on the root lattice. The elements w of the Weyl group W are made to act on monomials by

$$w(x^{\gamma})=x^{w(\gamma)}$$

and by linearity on all Laurent polynomials. For example, if $w(\alpha_1,\alpha_2)=(-\alpha_2,\alpha_1)$, then

$$w(x^{-1}y^2+3+x^5y^{-2})=x^{-2}y^{-1}+3+x^2y^5.$$

A Laurent polynomial G is *symmetric* with respect to the Weyl group W if $w(G)=G$ for every w in W. The *sign* of an element w of W, written $\operatorname{sgn}(w)$, may be defined as [C, p. 18] $(-1)^{n(w)}$, where $n(w)$ is the number of positive roots that w turns into negative roots, i.e., the number of elements in the set $w(R^+)\cap R^-$. A Laurent polynomial G is *antisymmetric* if for any w in the Weyl group W, $w(G)=\operatorname{sgn}(w)G$.

C.T. stands for "the constant term of" (in $x=(x_1,\cdots,x_l)$), and $|A|$ denotes the number of elements of the finite set A. The letter l usually denotes the rank of R, and $d_1,\cdots,d_l$, are the "fundamental invariants" [C, p. 155] of R.

The $(\ )_a$ q-notation will be used extensively. $(y;Q)_a$, the q-analogue of $(1-y)^a$ to base Q, is defined by

$$(y;Q)_a=(1-y)(1-Qy)(1-Q^2y)\cdots(1-Q^{a-1}y),$$

and whenever the "base" Q happens to be q we will omit it: $(y)_a=(y;q)_a$. The standard base of Euclidean space is denoted by $\{e_i\}$, $e_i=(0,0,\cdots,0,1,0,0,\cdots,0)$, where all the components are zero except the i component that is 1. Of course $x^{e_i}=x_i$.

1. Conjectures. In 1962, in his study of the statistical theory of complex systems, Dyson [D1] conjectured

$$\text{(D)}\qquad \text{constant term of } \prod_{1\le i\ne j\le n}\left(1-\frac{x_i}{x_j}\right)^a=\frac{(na)!}{a!^n}.$$

His conjecture was soon proved by Gunson [Gu] and Wilson [W] and Good [Goo] gave a beautiful proof some years later.

When Macdonald saw Dyson's conjecture (D) he saw the root system A_{n-1}. Indeed, since

$$A_{n-1}=\{e_i-e_j;\ 1\leqq i\neq j\leqq n\} \quad \text{and} \quad x^{e_i-e_j}=\frac{x_i}{x_j},$$

(D) can be written as

$$\text{constant term of } \prod_{\alpha\in A_{n-1}}(1-x^{\alpha})^{a}=\frac{(na)!}{a!^{n}}.$$

He then wondered what happens if A_{n-1} is replaced by other root systems.

The case $a=1$ of Dyson's conjecture (D) is an almost immediate consequence of the Vandermonde determinant identity and the constant term then is $n!=n(n-1)\cdots(2)$. Now the Vandermonde determinant identity has a celebrated root-system analogue: Weyl's denominator identity (e.g., [C, p. 149]), and imitating the argument that proved (D) for $a=1$ yields, for any root-system R with Weyl group W,

$$\text{C.T.} \prod_{\alpha\in R}(1-x^{\alpha})=|W|.$$

For $R=A_{n-1}$, $W=S_n$ and since $|W|=|S_n|=n!$, this agrees with the $a=1$ case of (D).

So the $a=1$ case of (D) has a nice root-system analogue. What about general a? It is well known that $|W|$ factorizes nicely [C, 9.3.4(i), p. 133]:

$$|W|=d_1d_2\cdots d_l,$$

where $d_1,\cdots,d_l$ are the "fundamental invariants" of the Weyl group W (these fundamental invariants are, among other things, the degrees of the generators of the algebra of polynomials invariant under W). For A_{n-1} these invariants are $2,3,\cdots,n$ (the degrees of the elementary symmetric functions!). Rewriting the right-hand side of (D) as

$$\binom{2a}{a}\cdots\binom{na}{a},$$

Macdonald [Ma3] conjectured that

$$\text{(M)} \qquad \text{constant term of } \prod_{\alpha\in R}(1-x^{\alpha})^{a}=\binom{d_1a}{a}\cdots\binom{d_la}{a}.$$

Macdonald was also able to prove the special case $a=2$, and by using Selberg's integral [Se] he proved the B_n, C_n, and D_n cases. Recently Habsieger [Hab1] and Zeilberger [Z2] proved the G_2 case. For $R=F_4$, E_6, E_7, and E_8, (M) is still open, as far as I know.

Next Macdonald went on to formulate a "q-analogue" of (M). Andrews ([An1]; see also [An2]) already formulated a q-analogue of (D) in 1975. Actually Andrews conjectured a q-analogue of a more general conjecture of Dyson, and his conjecture specializes to the following q-analogue of (D):

$$\text{(qD)} \qquad \text{C.T.} \prod_{1\leqq i<j\leqq n}\binom{x_i}{x_j}_k\binom{qx_j}{x_i}_k=\frac{[nk]!}{[k]!^{n}}\left(=\begin{bmatrix}2k\\k\end{bmatrix}\cdots\begin{bmatrix}nk\\k\end{bmatrix}\right).$$

The general Andrews conjecture was proved in [Z-B].

Motivated by this and (M) Macdonald [Ma3] conjectured

$$\text{(qM)} \qquad \text{C.T.} \prod_{\alpha\in R^{+}}(x^{\alpha})_k(qx^{-\alpha})_k=\begin{bmatrix}d_1k\\k\end{bmatrix}\cdots\begin{bmatrix}d_lk\\k\end{bmatrix}.$$

Macdonald [Ma3] was able to prove (qM) for $k=1, 2$ and $k=\infty$. For $k=\infty$ (qM) is a consequence of his own famous Macdonald Weyl identities [Ma2] (many special cases of which were known to Dyson [D2], but Dyson "missed the opportunity" to see the connection to root systems). For general k (qM) is only known to date for $R=A_n$ [Z-B] and G_2 ([Hab1], [Z2]). Hanlon [Han1] did the limiting case $n=\infty$ of B_n, C_n, and D_n.

One of the greatest delights of mathematics is the interplay between the abstract and the concrete, the general and the special. Whenever one has a general result or conjecture, it is very instructive to see what it says in special cases, and studying these special cases often sheds new light on the general case. Morris [Mo] took Macdonald's conjectures and made them explicit for all the root systems. Then by studying the G_2 case and playing with MACSYMA he was able to come up with a more general G_2-Macdonald conjecture, involving two parameters a and b instead of the single parameter k:

$$\text{C.T.} \prod_{\alpha \in \text{short } G_2} (1-x^{\alpha})^{a} \prod_{\alpha \in \text{long } G_2} (1-x^{\alpha})^{b} = \frac{(3a+3b)!(3b)!(2a)!(2b)!}{(2a+3b)!(a+2b)!(a+b)!a!b!b!}.$$

This was encouraging because it always helps to have more parameters (recall Polya's dictum: "the more general the easier"). Indeed Good's ([Goo]; see also [An2], [As3]) elegant proof of Dyson's conjecture (D) proceeds by proving the more general formula (also conjectured by Dyson [D1]):

$$\text{(D')} \qquad \text{C.T.} \prod_{1 \leq i \neq j \leq n} \left(1-\frac{x_i}{x_j}\right)^{a_i} = \frac{(a_1+\cdots+a_n)!}{a_1!\cdots a_n!},$$

and the extra elbow room provided by the n parameters $a_1, \cdots, a_n$ is crucial.

Morris sent his G_2 conjecture to Macdonald and, once again, Macdonald saw the right root-system generalization ([Ma3], [Mo]). Now there is a parameter associated with each root length. (Since A_l, D_l, E_6, E_7, E_8 have only one root length the generalization is void for them. For B_l, C_l, G_2, F_4 we have two parameters and BC_l has three parameters.)

Macdonald soon found a q-analogue [Ma3, 3.1]: if k_α are nonnegative integers such that $k_\alpha = k_\beta$ if α and β have the same length, then

$$\text{(qM-M1)} \qquad \text{C.T.} \prod_{\alpha \in R^+} (x^{\alpha})_{k_\alpha} (qx^{-\alpha})_{k_\alpha} = \text{a certain explicit product.}$$

I already mentioned that the case $k=\infty$ of (qM) is a consequence of Macdonald's Weyl identitities [Ma1]. These identities are the analogue of the Weyl denominator formula for affine root systems. (Incidentally these were "the tip of the iceberg" that motivated the representation theory of Kac-Moody algebras [Kac, p. xiii], but that's another story.) It turned out that the Macdonald-Morris conjectures (qM-M1) can be viewed as the "truncated form" of Macdonald's identities for the so-called $S(R)$ affine root systems ([Ma3, p. 999]; see [Ma1] and [Mo] for definitions of affine root systems). The classification theory of affine root systems [Ma1] says that the irreducible ones are either of the form $S(R)$ or $S(R)^{\vee}$. It was thus natural for Macdonald to formulate his conjectures as the truncated form of his identities and that led to the ultimate generalization ([Ma3, Conjecture 3.3], [Mo, pp. 25, 26]): Let k_α be as before and let u_α be certain *constant* integers (depending on the affine root system) that satisfy $u_\alpha = u_\beta$ whenever α and β have the same length (see [Ma1] or [Mo] for their values, for example for $S(G_2)^{\vee}$, $u_{\text{short}}=1$ and $u_{\text{long}}=3$). Let R be the underlying finite root system;

then,

(qM-M2) C.T. $\prod_{\alpha\in R^+}(x^\alpha; q^{u_\alpha})_{k_\alpha}(q^{u_\alpha}x^{-\alpha}; q^{u_\alpha})_{k_\alpha}$ = a certain explicit product.

The Macdonald conjectures, like most interesting mathematics, lie on the crossroads of several subjects, and so appeal to a wide spectrum of mathematicians. Lie algebraists suspect that, like the Macdonald identities, they are the tip of a deep algebraic iceberg [Han1]–[Han3], [Stanl1], [Stanl2]. Analysts [Mo], [As1]–[As3] see many interesting examples of multivariate hypergeometric series identities, "a topic about which little is currently known" [Mo, p. 4]. Geometers wonder whether there are things about root systems that they do not know, and combinatorialists [Z–B], [Br1], [Br2], [C-H], [B-G] are challenged to develop a combinatorial theory of Weyl groups that will emulate the rich theory of the symmetric group.

But regardless of our parochial interests and prejudices, we are all awed by the simplicity of these conjectures. The statement of the Macdonald conjectures, for any specific root system, can be explained to a high school student, but the proofs elude us.

2. Approaches. I will now give a very brief survey of the various approaches that have been used to tackle the Macdonald conjectures.

Selberg's integral. This fascinating generalization of Euler's beta integral was discovered by Selberg [Se] in 1944 but lay dormant for about 35 years, partially because it was ahead of its time, partially because it was written in Norwegian and partially because Selberg wrote it before he got *really* famous. This sleeping beauty awoke from its deep slumber when Enrico Bombieri consulted Selberg about a certain conjectured definite integral of Mehta [Me], [As3], [Ma3] and Selberg dug his old paper out of his files. It turned out that Mehta's conjecture (that has been open for about 15 years) is an easy consequence, via a limiting process, of Selberg's integral.

Mehta's conjecture [Me], which can be thought of as an integral analogue of Dyson's conjecture (D), also received root-system analogues by Macdonald [Ma3, § 5]. Beckner and Regev (see [Ma3, § 5]) showed how Selberg's integral can be used to get these root-system-Mehta conjectures for the classical root systems.

Macdonald [Ma3] showed, by a clever change of variable, that Selberg's integral is equivalent to the BC_n, $q=1$, case of (qM-M), which implies (M) for B_n, C_n, and D_n. Using a corollary of Selberg's integral, due to Morris [Mo, p. 94], Zeilberger [Z2], and Habsieger [Hab1] proved the G_2 case of (M). Aomoto [Ao] has recently found a very ingenious proof of Selberg's integral by using integration by parts, recurrences, and symmetry; see [As3] for a nice account.

By employing Jackson's q-analogue of integration, Askey [As1] formulated an elegant q-analogue of Selberg's integral that has recently been proved by Kadell [Kad1] and by Habsieger [Hab2]. Kadell q-analogized Aomoto's proof and Habsieger used Selberg's original method coupled with some brilliant ideas of his own. Kadell and Habsieger also showed that their Askey- q-Selberg identity implies the q-analogue of Morris' identity mentioned above. This q-analogue, conjecture by Morris himself [Mo], enabled Habsieger [Hab1] and Zeilberger [Z2] to prove the G_2 case of (qM-M1). Incidentally, Kadell and Habsieger's q-Morris identity contains, as a special case, the A_n case of (qM-M) (first proved in [Z-B]).

Counting tournaments. I already mentioned that the case $a=1$ of (D) follows from the Vandermonde determinant identity. The case $n=3$ is also classical, being equivalent to Dixon's identity [An1]. Both these classical identities received beautiful combinatorial proofs. Gessel [Ge] (see also [An2, 4.4]) gave an elegant graph-theoretical proof of Vandermonde's determinant identity by counting tournaments, and

Foata [F], [C-F] gave a gorgeous proof of Dixon's identity by using multitournaments on three players.

Combining these two pretty proofs, Zeilberger [Z1] managed to give a purely combinatorial proof of Dyson's conjecture (D′) (and thus of (D)). In that paper Zeilberger wrote: "We believe that our proof has a good chance of being generalized, because most combinatorial proofs involving binomial coefficients have q-analogues. However, another idea is still needed since the obvious q-generalization fails." The "obvious generalization" was to q-count words by using either the number of inversions or the major index as the "statistics" because both yield the q-multinomial coefficients. But neither of these worked. The new idea that was needed was to introduce a brand-new statistic, the z-index, and to prove that it, too, yields the q-multinomial coefficients. This was done in [Z-B], which contains a proof of Andrews' conjecture (and hence of the A_{n-1} case of (qM)).

Motivated by the success of the combinatorial method, there were attempts to extend it to general root systems [Br1], [Br2], [C-H]. Although these papers contain some very promising ideas, they failed, so far, even to prove the G_2 case. I should also mention [B-G], that, using the methods of [Z-B], contains interesting extensions of Andrews' conjecture, and [Gr], that gives an elegant MacMahon-style combinatorial proof for the above-mentioned fact that the z-statistics yield the q-multinomial coefficients.

Lie algebra cohomology. Hanlon [Han2], [Han3] found an interesting formulation and refinement of Macdonald's conjectures in the context of the cyclic homology of the exterior product of a Lie algebra with $\mathbb{C}[t, t^{-1}]$. Besides the considerable intrinsic merit of this approach, it also serves to make the conjectures accessible and appetitizing to all those sophisticates who are unwilling or unable to think in terms of the simple formulation of the original conjectures.

Hypergeometric $SU(N)$. Milne [Mi] found an elegant elementary proof of the $A_l^{(1)}$ case of Macdonald's identities. It is very possible that Milne's deep generalized hypergeometric theory will, one day, contain the Macdonald conjectures as a very special case.

Schur functions. Stanley [Stanl1], [Stanl2] found an interesting connection between the A_n cases and Schur functions. This connection was further explored by Stembridge [Ste1], [Ste2] and Goulden [Gou]. While these works do not try to prove the Macdonald conjecture per se, they Schur do give lots of insight. Indeed, it was exactly this study that led Stembridge [Ste3] to his elegant proof discussed below. It is not unlikely that a similar study of characters of general simple Lie algebras will lead us in the right direction.

I should also mention the interesting character sums analogues of Evans [E] and the fascinating connection between Mehta type integrals, PI rings and the representation of the symmetric group found by Regev [R1], [R2], and further explored by Cohen and Regev [C-R].

In April 1986, Dennis Stanton told me that John Stembridge had a short and elementary proof of the A_{n-1} case of (qM) (or equivalently, the equal parameter case of Andrews' q-Dyson conjecture). At first I was only mildly interested, since Kadell and Habsieger had just then completed, independently, the proof of Askey's conjectured q-Selberg integral ([Hab2], [Kad1], mentioned above) and also showed that it implies the q-Morris conjecture, that in turn implies the A_{n-1} case of (qM). In fact, I saw [Z3] how to use the Aomoto-Kadell method to get the q-Morris directly, without q-Selberg.

I wrote to John Stembridge anyway, requesting an account of the proof, and received from him a barely legible xerox copy of a three-page handwritten sketch that

Dennis Stanton had prepared. When I finally understood the proof I got excited. At long last a proof that has a "root-systemy" flavor! Although the proof was only of the A_{n-1} case, it had some universal root-system elements in it, and it used properties of the symmetric group that pass verbatim to any Weyl group, that I was sure that it should extend to the general case.

This was, of course, also realized by John Stembridge himself as he pointed out later when he finally got around to writing the paper [Ste3]. However, his proof took advantage of a certain "miracle" that seemed to occur only for A_{n-1}. Surely what was needed was to get rid of the dependence on the miracle, possibly by sacrificing elegance. I thought about that all through the summer (while taking care of my newborn daughter Tamar) and the result is this paper. (The fall was spent programming the algorithm and debugging the programs. If nothing else, this project made me a fairly competent *C* programmer.)

As John Stembridge told me himself, his proof, as well as parts of his impressive thesis [Ste1], were largely motivated and inspired by Dennis Stanton's ingenious proof of Macdonald's Weyl denominator identity for the classical root-systems [Stant1], [Stant2].

Using these beautiful ideas of Stembridge and Stanton, I will present a method that systematically handles the Macdonald conjectures for any given, fixed, root system, provided there are sufficient computer resources, and, for the time being, some luck. What I do know for sure is that it works for the (already known) A_2 and G_2 cases and for the (so far open) G_2^ν case (§ 9). Besides, I am almost sure that the element of luck can be disposed of and that the method can be proved to constitute an effective algorithm for settling the Macdonald and the Macdonald-Morris conjectures for any given root system. Of course, that by itself would not constitute a proof, or even an effective algorithm for the general conjecture, because there are an infinite number of root systems.

On the other hand, it is very possible that the A-D cases of the Macdonald and Macdonald-Morris conjectures will soon be settled by either using the Askey q-Selberg integral [Kad1], [Hab2] directly, or by using similar methods of proof. In that case we will only be left with the seven exceptional cases (G_2 and its dual, F_4 and its dual, and E_6, E_7, and E_8, but since the first two are already known this leaves us with five cases). These should succumb to the method of this paper (at least in principle, and barring very bad luck). But even if that would turn out to be the case, it would certainly not be the proof from *the book*. The ultimate proof should be "classification-free" and take care of all root systems at once.

To give a very apt analogy, the Weyl denominator formula [C, p. 149] can be proved case by case. A_{n-1} is just the Vandermonde determinant identity, which is an elementary exercise in determinants. The cases B_n, C_n, and D_n also specialize to simple algebraic identities that can be easily proved by induction. The remaining exceptional cases, G_2-E_8, give rise to finite polynomial identities that can be checked by computer (although I have to admit that, for E_8, even the CRAY will take "a while" to handle the 2^{120} terms). However, there is a beautiful "classification-free" proof of Weyl's identity that can be found in Carter's book ([C, § 10.1]).

I believe that besides the instant gratification that the present method brings, it is also an important step toward the ultimate proof. Unlike any previous approach, it makes use of the general root-system-Weyl group framework, and thus may pave the way to the final proof. In addition, it also provides a "laboratory" for computing other coefficients, besides the constant term, for any specific root system (see below). This may lead us to formulate a yet more general conjecture, and this more general conjecture

may very well turn out to be much easier to prove.

3. Antisymmetry.

No two fermions can exist in identical quantum states.
(*Wolfgang Pauli*)

Let R be a root system and let us define

$$F'_k(x)=\prod_{\alpha\in R^+}(x^\alpha)_k(qx^{-\alpha})_k, \qquad H'_k=\text{C.T. } F'_k(x). \tag{3.1}$$

Macdonald's conjecture (qM) asserts that H'_k has a nice explicit form (namely, the right-hand side of (qM)). In any case, whether (qM) is true or false, our goal will be to compute H'_k. It turns out (and this observation is due to Macdonald, although Stembridge was the one to realize its full significance) that one can consider instead

$$F_k(x)=\prod_{\alpha\in R^+}(x^\alpha)_k(qx^{-\alpha})_{k-1} \tag{3.2}$$

and

$$H_k=\text{C.T. } F_k(x).$$

This is so because of the fact, soon to proved, that H_k and H'_k are related by a simple formula

$$H'_k=H_k\prod_{i=1}^{l}\left(\frac{1-q^{kd_i}}{1-q^k}\right). \tag{3.3}$$

The reason why it is better to consider the constant term of F_k rather than that of F'_k is that F_k is a much nicer Laurent polynomial: it is almost antisymmetric.

Indeed, by peeling off the first layer out of the $(x^\alpha)_k$ in (3.1) we get (since $(y)_k=(1-y)(qy)_{k-1}$)

$$\begin{aligned}F_k(x)&=\prod_{\alpha\in R^+}(1-x^\alpha)\prod_{\alpha\in R^+}(qx^\alpha)_{k-1}(qx^{-\alpha})_{k-1}\\ &=\prod_{\alpha\in R^+}x^{\alpha/2}(x^{-\alpha/2}-x^{\alpha/2})\prod_{\alpha\in R}(qx^\alpha)_{k-1}\\ &=x^\delta\prod_{\alpha\in R^+}(x^{-\alpha/2}-x^{\alpha/2})\prod_{\alpha\in R}(qx^\alpha)_{k-1}\end{aligned} \tag{3.4}$$

(δ is one half the sum of all the positive roots). Let

$$G_k(x)=x^{-\delta}F_k(x). \tag{3.5}$$

Then, because of (3.4)

$$G_k(x)=\prod_{\alpha\in R^+}(x^{-\alpha/2}-x^{\alpha/2})\prod_{\alpha\in R}(qx^\alpha)_{k-1}. \tag{3.6}$$

We claim that $G_k(x)$ is antisymmetric. Indeed, the second product is symmetric because any element $w\in W$ sends R to itself [C, p. 13, line 4] and the first product is antisymmetric for the same reason, only now we get a minus sign whenever a positive root α is sent to a negative root (i.e., whenever $w(\alpha)\in R^-$). Thus the effect of applying $w\in W$ on the first product of (3.6) is to multiply it by $(-1)^{n(w)}$ where $n(w)=|w(R^+)\cap R^-|$ and this is equal to the sign of w [C, p. 18]. It thus follows that $G_k(x)$ itself is antisymmetric.

From now on, we will forget all about F_k (and certainly about F'_k) and work solely with G_k, noting that the quantity of interest, H_k, is given in terms of G_k by

$$H_k = \text{C.T.}\,(x^{\delta} G_k) C_{x^{-\delta}} G_k. \tag{3.7}$$

Our goal, to be pursued in the next two sections, is to find H_k. But we must show now, following Stembridge [Ste3], that H_k is indeed related to H'_k as promised by (3.3).

Indeed, since $(y)_k = (y)_{k-1}(1-q^k y)$, we have (by (3.1) and (3.2))

$$\begin{aligned} H'_k &= \text{C.T.} \prod_{\alpha \in R^+} (1-q^k x^{-\alpha}) F_k = \text{C.T.} \left[\prod_{\alpha \in R^+} (1-q^k x^{-\alpha}) \right] x^{\delta} G_k \\ &= \text{C.T.} \prod_{\alpha \in R^+} [(1-q^k x^{-\alpha}) x^{\alpha/2}] G_k = \text{C.T.} \prod_{\alpha \in R^+} (x^{\alpha/2} - q^k x^{-\alpha/2}) G_k. \end{aligned} \tag{3.8}$$

When the product on the extreme right is expanded we get $2^{|R^+|}$ terms, since each term in the product corresponds to a pair of opposite roots, and each term in the resulting huge sum corresponds to choosing, for every α in R^+, whether to take it or its negative. This prompts us to define a *choice set* Ω, as a subset of R such that for each $\alpha \in R^+$ either $\alpha \in \Omega$ or $-\alpha \in \Omega$.

We can now write the right-hand side of (3.8) as (set $t = q^k$),

$$\text{C.T.} \sum_{\Omega \text{ choice set}} (-t)^{|\Omega \cap R^-|} (x^{\text{sum}(\Omega)/2} G_k). \tag{3.9}$$

Here sum (Ω) denotes the (vector) sum of all the elements in Ω.

Now let us call a choice set a *bad guy*, if sum (Ω) lies on a reflecting hyperplane, i.e., there exists a root β such that $(\text{sum}\,(\Omega), \beta) = 0$. Otherwise let us call it a *good guy*. The sum in (3.9) can, of course, be written as

$$\text{C.T.} \sum_{\Omega \text{ good guy}} (-t)^{|\Omega \cap R^-|} (x^{\text{sum}(\Omega)/2} G_k) + \text{C.T.} \sum_{\Omega \text{ bad guy}} (-t)^{|\Omega \cap R^-|} (x^{\text{sum}(\Omega)/2} G_k). \tag{3.10}$$

The proof of (3.3) will continue *right after this.*

CRUCIAL LEMMA. *Let $G(x)$ be antisymmetric with respect to the Weyl group W and let γ be any vector of integers.*

(i) $\text{C.T.}\,(x^{w(\gamma)} G) = \text{sgn}\,(w)\ \text{C.T.}\,(x^{\gamma} G)$, *for each element w in the Weyl group W.*

(ii) *If γ lies on a reflecting hyperplane, i.e., there exists an $\alpha \in R$ such that $(\gamma, \alpha) = 0$, then* $\text{C.T.}\,(x^{\gamma} G) = 0$.

Proof of the Crucial Lemma.

Proof of (i).

$$\begin{aligned} \text{C.T.}\,(x^{w(\gamma)} G) &= \text{C.T.}\ w(x^{\gamma} w^{-1}(G)) \\ &= \text{C.T.}\, x^{\gamma} w^{-1}(G) = \text{C.T.}\, x^{\gamma}\ \text{sgn}\,(w^{-1}) G \\ &= \text{sgn}\,(w)\ \text{C.T.}\, x^{\gamma} G. \end{aligned}$$

In this chain of equalities we have used, in that order: (a) the definition of the action of w on a Laurent polynomial; (b) the fact that applying w on a Laurent polynomial never changes the consant term (because w is, among other things, a linear transformation, so $w(0) = 0$ and $w(x^0) = x^{w(0)} = x^0$); (c) the antisymmetry of G; (d) $\text{sgn}\,(w^{-1}) = \text{sgn}\,(w)$, and you can always take a constant out of C.T.

Proof of (ii). Let w_α be the Weyl reflection corresponding to the root α [C, p. 12]; then $w_\alpha(\gamma) = \gamma$ (since γ lies on the mirror that is perpendicular to α) and since $\text{sgn}\,(w_\alpha) = -1$, we have, by part (i),

$$\text{C.T.}\,(x^{\gamma} G) = \text{C.T.}\,(x^{w_\alpha(\gamma)} G) = (\text{sgn}\,(w_\alpha))\ \text{C.T.}\,(x^{\gamma} G) = -\text{C.T.}\,(x^{\gamma} G).$$

Thus $\text{C.T.}\,(x^{\gamma} G)$ is equal to its negative and must be zero. □

We now return to the proof of (3.3).

Because of part (ii) of the Crucial Lemma and the definition of a bad guy, the second sum in (3.10) vanishes. Now from Lemma 10.1.6. and its proof of [C, p. 147], or from Lemma 2.13 of [Ma2], it follows that if Ω is a good guy then there exists a w in the Weyl group W such that sum $(\Omega)/2 = w(\delta)$ and $\Omega = w(R^+)$. So a good choice set Ω uniquely determines $w \in W$ and vice versa. It is thus possible to write (3.9) as (note that $|w(R^+) \cap R^-| = n(w)$)

$$H'_k = \sum_{w \in W} (-1)^{n(w)} t^{n(w)} \text{ C.T. } (x^{w(\delta)} G_k).$$

But because of part (i) of the Crucial Lemma,

$$\text{C.T. } (x^{w(\delta)} G_k) = \text{sgn } (w) \text{ C.T. } (x^{\delta} G_k)$$

and since sgn $(w) = (-1)^{n(w)}$, we have that H'_k is equal to

$$H'_k = \left(\sum_{w \in W} t^{n(w)} \right) H_k,$$

and (3.3) follows because of the following beautiful identity due to Bott, Solomon, and Macdonald [Ma2] (see [C, p. 135 ff, p. 155])

$$\sum_{w \in W} t^{n(w)} = \prod_{i=1}^{l} \frac{1 - t^{d_i}}{1 - t}. \tag{3.11}$$

We should remark, though, that if one is only interested in one root system at a time (as we are in the present method), then we really do not need (3.11), since the left-hand side is just a specific polynomial that can be explicitly computed and, if desired, factorized.

4. Induction. This section constitutes my own twist on the Stembridge approach. Stembridge's [Ste3] inductive scheme, for A_n, was to creep along the coefficients of G_k (keeping k fixed) until one gets to a high enough coefficient whose value is equal to the H_k for A_{n-1}. So his induction was with respect to n, and his k stayed fixed. Our induction is with respect to k and the root system stays fixed.

Using $(y)_{k+1} = (1 - q^k y)(y)_k$, $(qy)_k = (1 - q^k y)(qy)_{k-1}$, (3.2) and (3.5), we have

$$H_{k+1} = \text{C.T. } (x^{\delta} G_{k+1}) = \text{C.T. } \left(x^{\delta} \prod_{\alpha \in R} (1 - q^k x^{\alpha}) G_k \right). \tag{4.1}$$

Now put $t = q^k$ and expand the product

$$x^{\delta} \prod_{\alpha \in R} (1 - t x^{\alpha}) = \sum_{\rho' \in S'} a_{\rho'}(t) x^{\rho'} \tag{4.2}$$

where S' is a certain *finite* set of vectors in the lattice generated by the roots and $a_{\rho'}(t)$ are polynomials in t. Now, each $\rho' \in S'$ is either on a reflecting hyperplane (a bad guy) or [C, Prop. 2.3.4, p. 22] there is a $w \in W$ and ρ in the fundamental chamber such that $\rho' = w(\rho)$. Thus defining S to be the set of all the W images of S' that lie in the fundamental chamber, the right-hand side of (4.2) can be written as

$$\sum_{\rho' bad} a_{\rho'}(t) x^{\rho'} + \sum_{\rho \in S} \sum_{w \in W} a_{\rho, w}(t) x^{w(\rho)} \tag{4.3}$$

where $a_{\rho, w}(t)$ are certain (easily computable) polynomials in t (some of which may be zero).

Substituting this into the right-hand side of (4.1) we get that the contribution from the first sum in (4.3) is zero (Crucial Lemma (ii)), and it follows from part (i) of the Crucial Lemma that

$$H_{k+1} = \sum_{\rho \in S} \sum_{w \in W} a_{\rho,w}(t)\ \text{C.T.}\ (x^{w(\rho)} G_k) = \sum_{\rho \in S} \left(\sum_{w \in W} a_{\rho,w}(t)\ \text{sgn}\ (w) \right) \text{C.T.}\ (x^{\rho} G_k). \tag{4.4}$$

Now for each $\rho \in S$, let

$$A_\rho(t) = \sum_{w \in W} a_{\rho,w}(t)\ \text{sgn}\ (w). \tag{4.5}$$

$A_\rho(t)$ is a certain explicitly computable polynomial in t. Going back to (4.4) we have

$$H_{k+1} = \sum_{\rho \in S} A_\rho(t)\ \text{C.T.}\ (x^{\rho} G_k). \tag{4.6}$$

One of the summands here is $\rho = \delta$, so we have expressed H_{k+1} in terms of C.T. $(x^{\delta} G_k) = H_k$ and a certain *finite* number of "neighboring coefficients." We have thus encountered the notorious "problem of uninvited guests" that crops up so often when trying to prove something by induction. One way out of this, the polite way, is to put up with these undesirable terms and conjecture that they too, have a certain explicit form, and then redo (4.6) to account for these as well (and cross our fingers that they will not bring in more undesirable terms). I do not see how to do it (at least not yet). The other way is the rude way. Get rid of these undesirable terms by expressing all of them in terms of the only term that we really care about: the one and only H_k.

5. Equations. This section will describe Stembridge's variation on an old trick in q-series, adapted to our needs. This trick converts a q-product in one variable $f(x)$ into a sum by computing $f(qx)/f(x)$. If this turns out to be a rational function, then cross-multiplying yields a functional equation relating $f(x)$ and $f(qx)$. By expanding $f(x)$ in a power series, this translates into a linear recurrence in the coefficients, that sometimes can be solved explicitly. However, attempting to use this method for multivariate products always produces a mess, unless we have antisymmetry on our side, and even then one has to be very careful.

So let us go to business. Using the definitions (3.2) and (3.5), we have

$$G_k(x) = x^{-\delta} \prod_{\alpha \in R^+} (x^{\alpha})_k (qx^{-\alpha})_{k-1}. \tag{5.1}$$

Recall that $x = (x_1, \cdots, x_l)$, $\alpha = (\alpha_1, \cdots, \alpha_l)$ and $x^{\alpha} = x_1^{\alpha_1} \cdots x_l^{\alpha_l}$. Define

$$f_\alpha(x) = (x^{\alpha})_k (qx^{-\alpha})_{k-1};$$

then if $\alpha_1 = 0$, $f_\alpha(x_1 \leftarrow qx_1) = f_\alpha(x)$, and in general (we assume, without loss of generality (see Introduction) that α has integer coordinates)

$$\frac{f_\alpha(x_1 \leftarrow qx_1)}{f_\alpha(x)} = \frac{(q^{\alpha_1} x^{\alpha})_k (q^{1-\alpha_1} x^{-\alpha})_{k-1}}{(x^{\alpha})_k (qx^{-\alpha})_{k-1}}. \tag{5.2}$$

Now by making all the $(\ )_k$ explicit and using telescoping, we easily obtain

$$\frac{f_\alpha(x_1 \leftarrow qx_1)}{f_\alpha(x_1)} = \frac{p_\alpha(x)}{q_\alpha(x)} \tag{5.3}$$

where, if $\alpha_1 > 0$,

$$\begin{aligned} p_\alpha(x) &= (1 - q^k x^{\alpha}) \cdots (1 - q^{k+\alpha_1 - 1} x^{\alpha}), \\ q_\alpha &= (q^{k-1} - x^{\alpha}) \cdots (q^{k-1} - q^{\alpha_1 - 1} x^{\alpha}), \end{aligned} \tag{5.4a}$$

and if $\alpha_1 < 0$,

$$\begin{aligned} p_\alpha(x) &= (q^{k-1} - q^{-1}x^\alpha) \cdots (q^{k-1} - q^{\alpha_1}x^\alpha), \\ q_\alpha(x) &= (1 - q^{k-1}x^\alpha) \cdots (1 - q^{k+\alpha_1}x^\alpha). \end{aligned} \tag{5.4b}$$

Since for all root-systems [C, pp. 47-49] $-2 \leqq \alpha_1 \leqq 2$, p_α, q_α, are at worst quadratic in x^α.

It follows from (5.3) and the definition (5.1) that

$$\frac{G_k(x_1 \leftarrow qx_1)}{G_k(x_1)} = q^{-\delta_1} \prod_{\alpha \in R^+} \frac{p_\alpha(x)}{q_\alpha(x)} = \frac{P(x, q^k, q)}{Q(x, q^k, q)}, \tag{5.5}$$

say, where P and Q are certain, explicitly computable, polynomials in $x = (x_1, \cdots, x_l)$, q^k and q.

Now, by cross-multiplying (5.5), we get the functional equation

$$Q(x)G_k(x_1 \leftarrow qx_1) = P(x)G_k(x). \tag{5.6}$$

Out of this functional equation we can get many linear equations relating various coefficients of G_k. For any vector β in the lattice generated by R, we will get a linear equation E_β, involving coefficients C.T. $(x^\gamma G_k)$ for γ in a certain set of vector exponents $Ex(\beta)$, that is contained in the fundamental chamber.

The way to do this is to first multiply both sides of (5.6) by x^β and then apply the functional C.T.

$$\text{C.T.}\,[x^\beta Q(x)G_k(x_1 \leftarrow qx_1)] = \text{C.T.}\,[x^\beta P(x)G_k(x)]. \tag{5.7}$$

We now plug into (5.7) the expanded form of P and Q (remember that P and Q are certain explicit polynomials that we have to compute in order to perform the algorithm). Then we use the linearity of C.T. and get on the right-hand side a linear combination of creatures of the form C.T. $[x^\gamma G_k]$. On the left-hand side we get a linear combination of entities of the form C.T. $[x^\gamma G_k(x_1 \leftarrow qx_1)]$. These should be converted to the previous form using the obvious relation

$$\text{C.T.}\,[x^\gamma G_k(x_1 \leftarrow qx_1)] = q^{-\gamma_1}\,\text{C.T.}\,[x^\gamma G_k]. \tag{5.8}$$

We now use the Crucial Lemma, discarding all the "bad" γ, i.e., those that are orthogonal to a root, and for any good γ' that is not in the fundamental chamber we find the unique $w \in W$ and γ in the fundamental chamber such that $\gamma' = w(\gamma)$ and rewrite C.T. $[x^{\gamma'} G_k]$ as sgn (w) C.T. $[x^\gamma G_k(x)]$. Then we collect all the terms and bring them to the left-hand side and get a certain linear equation

$$E_\beta\colon \quad \sum a_\gamma(q^k, q)\,\text{C.T.}\,(x^\gamma G_k) = 0 \tag{5.9}$$

where the sum is over a *finite* set $Ex(\beta)$ of exponents γ that lie in the fundamental chamber.

By a judicious choice of β we would hopefully obtain equations that only involve those $\rho \in S$ that feature in (4.6). Hopefully there would be $|S| - 1$ such independent equations. (Of course it would also be all right if we could say the same thing about some set that contains S.) By a proper choice of β it is always possible to get an equation that involves C.T. $(x^\delta G_k) = H_k$.

Solving this system of $|S| - 1$ homogeneous equations, at least one of which involves H_k, we should be able to express all the unknowns as H_k times some rational function in q^k and q. This is so since the coefficients in the system are polynomials in q^k and q. We have thus found explicit expressions for all the terms that feature in (4.6) in

terms of H_k, and plugging them in we will get H_{k+1}/H_k, a certain rational function in q^k and q. Calling the conjectured value of H_k by the name of R_k (R_k is the right-hand side of (qM) divided by (3.11)), we can then compare H_{k+1}/H_k with R_{k+1}/R_k. Since obviously $H_1 = R_1$, the fate of (qM) will be determined by whether or not H_{k+1}/H_k is equal to R_{k+1}/R_k.

6. Implementation. This can be, and has been, implemented on a computer. The input is the root system R, and it is necessary to know the Weyl group W (this is given in the *planches* of [Bo]). It is very easy to write a routine to check whether a given vector is a *bad guy* (just do-loop the inner product along R^+). Then you need to write a Weyl-sorting routine that given any good vector in the root-lattice finds its image in the fundamental chamber and the sign of the element w in W that sends it there. Of course you also need a polynomial multiplication routine (which you can easily jot down yourself, no need for MACSYMA). This is enough to produce (4.6) and the $P(x)$ and $Q(x)$ of (5.5).

Now comes the creative part, experimenting with various β's that will give an equation E_β that involves the relevant coefficients that feature in (4.6). For those root systems for which $-1 \leqq \alpha_1 \leqq 1$ (most of them) the choice $\beta = -\delta$ will produce a tautology: $0 = 0$, because the only survivor, after applying part (ii) of the Crucial Lemma, is C.T. $[x^\delta G_k] = H_k$. It is thus likely that for β near $-\delta$ we will get relatively few terms.

Once you have $|S| - 1$ independent equations you solve them and plug the solutions into (4.6). You will never have to see (or print out) the solutions of the system (5.9), because it can all be done internally (in MACSYMA this amounts to finishing your lines with dollar signs rather than with semicolons). You will not even have to see or print out the resulting rational function H_{k+1}/H_k obtained by plugging in the solutions of the system (5.9) into (4.6).

All you have to do is enter the rational function R_{k+1}/R_k (you can even write a routine for that) and ask the computer to output the difference between these two rational functions. If you get ZERO then you have proved (qM) for your particular root system. If you get something else then you have *disproved* (qM). Either that or (more likely), you have made an error somewhere.

7. A_2. The new method will now be illustrated on the simplest nontrivial case, the root system A_2. Of course this case is already well known, even classical (it is equivalent to Jackson's q-Dixon identity [An1]), and the proof that we present here is perhaps the longest and ugliest ever. But in order to learn how to use machine guns to kill elephants one should first practise on flies. Another reason for doing the A_2 case is that its results will be needed in § 9, when we do G_2^ν, and this will make the paper self-contained. The present example is simple enough that it can be done by hand, and the reader is encouraged to check all the steps and to supply all the details.

Equation (qM) says, in its equivalent formulation derived in § 3, that if

$$F_k = \left(\frac{x_1}{x_2}\right)_k \left(\frac{x_1}{x_3}\right)_k \left(\frac{x_2}{x_3}\right)_k \left(q\frac{x_2}{x_1}\right)_{k-1} \left(q\frac{x_3}{x_1}\right)_{k-1} \left(q\frac{x_3}{x_2}\right)_{k-1},$$

$$H_k = \text{C.T. } F_k$$

and

$$R_k = \frac{(q)_{3k-1}}{(q)_{k-1}^2 (q)_k (1-q^{2k})},$$

then $H_k = R_k$.

To get R_k from $R'_k=(q)_{3k}/(q)_k^3$ we used (3.3) with the fundamental invariants 2, 3 of A_2. A list of the fundamental invariants for all finite irreducible root systems can be found for example in the excellent appendices of [Bo], as well as in [C, p. 155].

Now a routine calculation shows that ($t=q^k$),

$$\frac{R_{k+1}}{R_k}=\frac{(1+t+t^2)(1-qt^3)(1-q^2t^3)(1+t)}{(1-qt)(1-q^2t^2)}. \tag{7.1}$$

Now it is easily checked that $R_1=1$ and $H_1=1$, so all we have to do is verify that H_{k+1}/H_k is equal to R_{k+1}/R_k. So let us compute H_{k+1}/H_k.

For A_2 we have (e.g., [Bo, p. 250] or [C, p. 46])

$$A_2^+=\{(1,-1,0);(1,0,-1);(0,1,-1)\}.$$

$\delta=(1,0,-1)$, and the Weyl group W is S_3, the symmetric group on three elements that acts by permuting the coordinates of $(\gamma_1,\gamma_2,\gamma_3)$ for γ in the root lattice. The bad guys are those vectors that have two of their coordinates equal.

Now we do (4.2), namely we expand

$$\frac{x_1}{x_3}\left(1-t\frac{x_1}{x_2}\right)\left(1-t\frac{x_2}{x_1}\right)\left(1-t\frac{x_1}{x_3}\right)\left(1-t\frac{x_3}{x_1}\right)\left(1-t\frac{x_2}{x_3}\right)\left(1-t\frac{x_3}{x_2}\right).$$

Discarding the bad guys, grouping the good guys into orbits under S_3, as in (4.3), plugging into (4.1) and using the Crucial Lemma yields, like in (4.4)-(4.6) (set $A(\rho)=\text{C.T.}[x^\rho G_k]$),

$$\begin{aligned}H_{k+1}=&(1+2t+3t^2+3t^3+3t^4+2t^5+t^6)A(1,0,-1)\\&-(t+t^2+2t^3+t^4+t^5)A(2,0,-2)\\&+(t^2+t^3+t^4)A(2,1,-3)+(t^2+t^3+t^4)A(3,-1,-2)-t^3A(3,0,-3).\end{aligned} \tag{7.2}$$

Thus, $S=\{(1,0,-1),(2,0,-2),(2,1,-3),(3,-1,-2),(3,0,-3)\}$, and we need to find four independent equations relating $\{A(\rho);\rho\in S\}$.

Now (5.3) becomes

$$\frac{f_{1-10}(qx_1,x_2,x_3)}{f_{1-10}(x_1,x_2,x_3)}=\frac{\left(1-q^k\frac{x_1}{x_2}\right)}{\left(q^{k-1}-\frac{x_1}{x_2}\right)},\qquad \frac{f_{10-1}(qx_1,x_2,x_3)}{f_{10-1}(x_1,x_2,x_3)}=\frac{\left(1-q^k\frac{x_1}{x_3}\right)}{\left(q^{k-1}-\frac{x_1}{x_3}\right)},$$

and (5.5) becomes ($\delta_1=1$),

$$\frac{G_k(qx_1,x_2,x_3)}{G_k(x_1,x_2,x_3)}=q^{-1}\frac{\left(1-q^k\frac{x_1}{x_2}\right)\left(1-q^k\frac{x_1}{x_3}\right)}{\left(q^{k-1}-\frac{x_1}{x_2}\right)\left(q^{k-1}-\frac{x_1}{x_3}\right)},$$

and (5.6) becomes

$$q\left(q^{k-1}-\frac{x_1}{x_2}\right)\left(q^{k-1}-\frac{x_1}{x_3}\right)G_k(qx_1,x_2,x_3)=\left(1-q^k\frac{x_1}{x_2}\right)\left(1-q^k\frac{x_1}{x_3}\right)G_k(x_1,x_2,x_3)$$

and multiplying out yields

$$\begin{aligned}&\left(q^{2k-1}-q^k\frac{x_1}{x_2}-q^k\frac{x_1}{x_3}+q\frac{x_1^2}{x_2x_3}\right)G_k(qx_1,x_2,x_3)\\&\quad=\left(1-q^k\frac{x_1}{x_2}-q^k\frac{x_1}{x_3}+q^{2k}\frac{x_1^2}{x_2x_3}\right)G_k(x_1,x_2,x_3).\end{aligned} \tag{7.3}$$

Experimenting with various β yields that $(0,1,-1)$, $(1,0,-1)$, $(0,2,-2)$, and $(1,1,-2)$ produce the desired equations (of course there are many other choices of β that will do). For each of these β, multiplying both sides of (7.3) by x^{β}, using (5.8) and the Crucial Lemma yields the equations:

$$E_{(0,1,-1)}\colon (1-q^{2k-1}-q^{k-1}+q^{k})A(1,0,-1)+(q^{-1}-q^{2k})A(2,0,-2)=0,$$

$$E_{(1,0,-1)}\colon (1-q^{2k-2})A(1,0,-1)+(q^{k-2}-q^{k})A(2,0,-2)+(q^{2k}-q^{-2})A(3,-1,-2)=0,$$

$$E_{(0,2,-2)}\colon (1-q^{2k-1})A(2,0,-2)+(q^{-1}+q^{k-1}-q^{k}-q^{2k})A(2,1,-3)=0,$$

$$E_{(1,1,-2)}\colon (q^{k-2}-q^{k})A(2,0,-2)+(q^{k-2}-q^{k})A(2,1,-3)+(q^{2k}-q^{-2})A(3,0,-3)=0.$$

Solving this system we get $(t=q^{k})$ (recall that $A(1,0,-1)=H_k$),

$$A(2,0,-2)=\frac{(t-q)(1-t^2)}{(1-t)(1-qt^2)}H_k,$$

$$A(3,-1,-2)=A(2,1,-3)=\frac{(q-t)(q-t^2)}{(1-qt)(1-qt^2)}H_k, \tag{7.4}$$

$$A(3,0,-3)=\frac{-t(1-q^2)(q-t)(1-q)(1-t^3)}{(1-qt)(1-qt^2)(1-q^2t^2)(1-t)}H_k.$$

This much was done by hand. Now using MACSYMA we can plug it all into (7.2) and get H_{k+1}/H_k. Then we subtract it from R_{k+1}/R_k given in (7.1). The answer is indeed zero and we have just proved (qM) for A_2.

Now that we know that H_k is indeed equal to what it is supposed to be, namely to R_k, we can plug that expression into (7.4) and get as a lagnappe explicit expressions for $A(2,0,-2)=A_k(2,0,-2)=\text{C.T.}[x_1^2x_3^{-2}G_k]=\text{C.T.}[x_1x_3^{-1}F_k]$, etc. This will be needed in § 9.

8. Modifications. Our method can be easily adapted to the more refined Macdonald–Morris conjectures (qM-M1) and (qM-M2). In fact, because of the added parameter it is even computationally faster. We will only treat (qM-M2), since (qM-M1) is just a special case of (qM-M2) ((qM-M1) corresponds to the $S(R)$ cases for which it is well known [Ma1], [Mo] that $u_\alpha\equiv 1$). It is also well known (for example from the classification theorem for finite root systems [Bo], [C] that all the irreducible reduced finite root systems have either just one root length (A_n, D_n, E_6, E_7, and E_8) or two root lengths (B_n, C_n, G_2, and F_4). The only nonreduced irreducible finite root systems, BC_n, have three different root lengths. Since (qM-M2) reduces to (qM) for all the single-length root systems, we will assume that the root systems have two root lengths, short and long, and leave it to the reader to do the appropriate obvious modifications for BC_n.

So let us rewrite (qM-M2) for two-lengths root systems. Denoting k_{short} by a, k_{long} by b, u_{short} by u_s, and u_{long} by u_l, we have

$$\text{(qM-M2')}\quad \text{C.T.}\left\{\prod_{\alpha\in R^+_{\text{short}}}(x^{\alpha};q^{u_s})_a(q^{u_s}x^{-\alpha};q^{u_s})_a\prod_{\alpha\in R^+_{\text{long}}}(x^{\alpha};q^{u_l})_b(q^{u_l}x^{-\alpha};q^{u_l})_b\right\}$$

$$=\text{a certain explicit product.}$$

The right-hand side can be looked up in [Ma3] or [Mo, pp. 25–26]. Its exact form is irrelevant for the purposes of the present method whose modest aim is to prove (qM-M2) for one root-system *at a time*, and as such does not care to look at the general pattern. Besides, the method should be able to compute the constant term in question *from scratch* and it is dishonest to "peek at the answer." In any case, for any specific root-system, it is possible to look up the explicit conjectured right-hand side from [Mo].

So let us call the polynomial inside the braces of (qM-M2′) $F'_{a,b}(x)$. We are interested in evaluating

$$H'_{a,b} = \text{C.T. } F'_{a,b}(x). \tag{8.1}$$

In analogy with § 3, we will consider instead

$$F_{a,b}(x) = \prod_{\alpha \in R^+_{\text{short}}} (x^\alpha; q^{u_s})_a (q^{u_s} x^{-\alpha}; q^{u_s})_{a-1} \prod_{\alpha \in R^+_{\text{long}}} (x^\alpha; q^{u_l})_b (q^{u_l} x^{-\alpha}; q^{u_l})_{b-1},$$

$$H_{a,b} = \text{C.T. } F_{a,b}(x).$$

Since the Weyl group W acts separately on the long roots and the short roots (as is obvious from the fact that the elements of W are isometries), the calculation of (3.3) can be carried verbatim to show that

$$G_{a,b}(x) =: x^{-\delta} F_{a,b}(x) \tag{8.2}$$

is antisymmetric.

For w in the Weyl group W let $n_s(w) = |w(R^+_{\text{short}}) \cap R^-|$ and $n_l(w) = |w(R^+_{\text{long}}) \cap R^-|$ (so $n(w) = n_s(w) + n_l(w)$). Define

$$W(t, s) = \sum_{w \in W} t^{n_s(w)} s^{n_l(w)}, \tag{8.3}$$

which for a fixed root system (and therefore a fixed Weyl group) is a specific polynomial. Macdonald [Ma2] has a wonderful formula for $W(t, s)$ as a product that is indexed over the positive roots (for $t = s$ it reduces to (3.11)), but it is not really needed for our present narrow-minded purposes.

A completely analogous argument to that of § 3 (only now we keep track of the short and long roots separately, with their respective parameters t and s) yields

$$H'_{a,b} = H_{a,b} W(q^{au_s}, q^{bu_l}). \tag{8.4}$$

We now want to evaluate

$$H_{a,b} = \text{C.T. } [x^\delta G_{a,b}].$$

The difference now is that we have two parameters a and b rather than the single parameter k. The induction step is similar to that described in § 4 only now we induct with respect to either a *or* b (I prefer a). Unlike the previous case where the base case was trivial, now $a = 1$ is no longer trivial but is essentially the (qM) conjecture for the subroot system consisting of the long roots

$$H_{1,b} = \text{C.T. } F_{1,b}(x)$$

where

$$F_{1,b}(x) = \prod_{\alpha \in R^+_{\text{short}}} (1 - x^\alpha) \prod_{\alpha \in R^+_{\text{long}}} (x^\alpha; q^{u_l})_b (q^{u_l} x^{-\alpha}; q^{u_l})_{b-1}. \tag{8.5}$$

Expanding

$$\prod_{\alpha\in R^+_{\text{short}}}(1-x^\alpha),$$

we can express $H_{1,b}$ as a certain linear combination of various coefficients of

$$\hat{F}_b(x):=\prod_{\alpha\in R^+_{\text{long}}}(x^\alpha;q^{u_l})_b(q^{u_l}x^{-\alpha};q^{u_l})_{b-1}.$$

Thus before we can embark on (qM-M2) for R we must do first (qM) for the subroot system R_{long} and find not only the constant term of $\hat{F}_b(x)$ but also some neighboring coefficients. It can be easily shown that these coefficients are among those "lagnappes" that we got anyway in system (5.9). For example, the long roots of G_2 constitute the root system A_2, and when we do G_2^ν in the next section we will use the A_2 information obtained in § 7. Similarly, before we can do F_4 we must do D_4, etc.

Having established the base case $a=1$, § 4 passes almost verbatim: (4.1) becomes

$$H_{a+1,b}=\text{C.T.}\,[x^\delta G_{a+1,b}(x)]=\text{C.T.}\left[x^\delta\prod_{\alpha\in R_{\text{short}}}(1-q^{au_s}x^\alpha)G_{a,b}\right] \tag{8.6}$$

and (4.6) becomes ($t=q^{au_s}$)

$$H_{a+1,b}=\sum_{\rho\in S}A_\rho(t)\,\text{C.T.}\,[x^\rho G_{a,b}]. \tag{8.7}$$

Now comes the analogue of § 5. We have to be a little careful because $G_{a,b}(x_1\leftarrow qx_1)/G_{a,b}(x)$ may not be a rational function. Instead we look for vectors of integers $z=(z_1,\cdots,z_n)$ such that

$$\frac{G_k(q^{z_1}x_1,\cdots,q^{z_n}x_n)}{G_k(x_1,\cdots,x_n)} \tag{8.8}$$

is a rational function. This can be achieved if u_s divides (α,z) for every short root α, and u_l divides (α,z) for every long α. Of course we will try to choose z in such a way that the rational function (8.8) is as simple as possible (in the next section $z=(2,1,0)$).

In analogy with § 5 we define

$$f_\alpha(x)=(x^\alpha;q^{u_\alpha})_{k_\alpha}(q^{u_\alpha}x^{-\alpha};q^{u_\alpha})_{k_\alpha-1}$$

where $k_\alpha=a$, $u_\alpha=u_s$ if α is short and $k_\alpha=b$ and $u_\alpha=u_l$ if α is long.

In analogy with (5.2) we have

$$\frac{f_\alpha(q^{z_1}x_1,\cdots,q^{z_n}x_n)}{f_\alpha(x_1,\cdots,x_n)}=\frac{(q^{(z,\alpha)}x^\alpha;q^{u_\alpha})_{k_\alpha}(q^{u_\alpha-(z,\alpha)}x^{-\alpha};q^{u_\alpha})_{k_\alpha-1}}{(x^\alpha;q^{u_\alpha})_{k_\alpha}(q^{u_\alpha}x^{-\alpha};q^{u_\alpha})_{k_\alpha-1}}.$$

So if we replace

$$q\leftarrow q^{u_\alpha},\qquad \alpha_1\leftarrow(z,\alpha)/u_\alpha,\qquad k\leftarrow k_\alpha$$

then (5.3) and (5.4) are still true. Equation (5.5) now becomes

$$\frac{G_{a,b}(q^{z_1}x_1,\cdots,q^{z_n}x_n)}{G_{a,b}(x_1,\cdots,x_n)}=q^{-(\delta,z)}\prod_{\alpha\in R^+}\frac{p_\alpha(x)}{q_\alpha(x)}=\frac{P}{Q} \tag{8.9}$$

where P and Q are explicitly computable polynomials in x, q, t, and s where $t=q^{au_s}$; $s=q^{bu_l}$.

Equations (5.6) and (5.7) are still true but with $G_k(x_1 \leftarrow qx_1)$ replaced by $G_k(q^{z_1}x_1, \cdots, q^{z_n}x_n)$ and instead of (5.8) we need

$$\text{C.T.}[x^\gamma G_{a,b}(q^{z_1}x_1, \cdots, q^{z_n}x_n)] = q^{-(\gamma,z)}\,\text{C.T.}[x^\gamma G_{a,b}]. \tag{8.10}$$

This follows from

$$\text{C.T.}[x^\gamma G_{a,b}(q^{z_1}x_1, \cdots, q^{z_n}x_n)]$$
$$= q^{-(\gamma,z)}\,\text{C.T.}[(q^{z_1}x_1)^{\gamma_1}\cdots(q^{z_n}x_n)^{\gamma_n}\cdots G_{a,b}(q^{z_1}x_1, \cdots, q^{z_n}x_n)]$$

and the fact that C.T. is unaffected by scaling.

Everything is as before; the only difference is that in (5.9) the coefficients a_γ depend on (t, s, q) where $t = q^{au_s}$ and $s = q^{bu_l}$, i.e.,

$$E_\beta\colon \quad \sum a_\gamma(t, s, q)\,\text{C.T.}[x^\gamma G_{a,b}] = 0. \tag{8.11}$$

Everything else translates smoothly. Solving the system we will express all the coefficients that feature in (8.7) as certain rational functions in (q, t, s) times $\text{C.T.}[x^\gamma G_{a,b}] = H_{a,b}$. Substituting the solutions thus obtained into (8.7) will give us the rational function $H_{a+1,b}/H_{a,b}$. Since we already have a formula for $H_{1,b}$ this easily yields a formula for $H_{a,b}$. Alternatively, if we believe that the conjectured value for $H_{a,b}$, let us call it $R_{a,b}$, has a good chance of being correct then all we have to do is look up $R'_{a,b}$ (the conjectured right-hand side of (qM-M2)) in [Mo] and then compute $R_{a,b}$ by dividing $R'_{a,b}$ by $W(q^{au_s}, q^{bu_l})$ of (8.3). We then compute $R_{a+1,b}/R_{a,b}$ (a rational function in (q, t, s)). Assuming that we have already checked that $H_{1,b} = R_{1,b}$, the status of the conjecture (qM-M2) for the particular root-system in question is determined by whether or not $H_{a+1,b}/H_{a,b} - R_{a+1,b}/R_{a,b}$ is zero or not.

9. G_2^ν.

THEOREM (G_2^ν case of (qM-M2)). *The constant term of*

$$F'_{a,b}(x, y, z) := \left(\frac{x}{y}; q\right)_a \left(\frac{z}{y}; q\right)_a \left(\frac{z}{x}; q\right)_a \left(\frac{z^2}{xy}; q^3\right)_b \left(\frac{xz}{y^2}; q^3\right)_b \left(\frac{yz}{x^2}; q^3\right)_b$$
$$\cdot \left(q\frac{y}{x}; q\right)_a \left(q\frac{y}{z}; q\right)_a \left(q\frac{x}{z}; q\right)_a \left(q^3\frac{xy}{z^2}; q^3\right)_b \left(q^3\frac{y^2}{xz}; q^3\right)_b \left(q^3\frac{x^2}{yz}; q^3\right)_b$$

is equal to

$$R'_{a,b} := \frac{(q;q)_{3a+3b}(q;q)_{3b}(q;q)_{2a}(q^3;q^3)_{a+3b}(q^3;q^3)_{2b}(q^3;q^3)_a}{(q;q)_{2a+3b}(q;q)_{a+3b}(q;q)_a^2(q^3;q^3)_{a+2b}(q^3;q^3)_{a+b}(q^3;q^3)_b^2}.$$

In this explict form the conjecture appears in Morris' thesis [Mo, p. 139]. It is alluded to in [As3, § 5, fifth sentence] and is mentioned explicitly in [As4].

From [Bo, pp. 274–275] or [Mo] or [C], $G_2^+ = \{(1, -1, 0), (0, -1, 1), (-1, 0, 1), (-1, -1, 2), (1, -2, 1), (-2, 1, 1)\}$; $\delta = (-1, -2, 3)$, and the Weyl group is the dihedral group of order 12, that is the direct product of S_3 with $\{I, -I\}$, where I denotes the identity mapping and $-I(\alpha, \beta, \gamma) = (-\alpha, -\beta, -\gamma)$. (It is a very instructive exericise for you to obtain the Weyl group yourself.) The bad guys are the vectors of integers $(\alpha_1, \alpha_2, \alpha_3)$ in which two coordinates are equal (those that are orthogonal to one of the short roots) and those vectors of integers in which one component is zero (those orthogonal to one of the long roots; recall that for all vectors in the root-lattice the sum of the components is zero).

A direct calculation shows that $W(t, s)$ of (8.3) is given by

$$W(t, s) = 1 + t + s + 2ts + ts^2 + t^2s + 2t^2s^2 + t^3s^2 + t^2s^3 + t^3s^3. \tag{9.1}$$

(For example -231 sends $(1,-1,0)$ to $(1,0,-1)$, a negative root; $(0,-1,1)$ goes to $(1,-1,0)$, a positive root; $(-1,0,1)$ goes to $(0,-1,1)$ a positive root; so $n_s(-231) = |(-231)(R^+) \cap R^-| = |\{(1,0,-1)\}| = 1$. Similarly, $n_l(-231) = 1$ and so -231 gives a contribution of ts to the sum of (8.3).)

$W(t, s)$ factorizes nicely, namely

$$W(t,s) = (1+t)(1+s)(1+ts+t^2s^2) \tag{9.2}$$

(compare [Ma2, p. 168]).

So with the notation of § 8 it follows from (8.4) that ($u_s = 1$, $u_l = 3$, $t = q^a$, $s = q^{3b}$)

$$H_{a,b} = \frac{H'_{a,b}}{(1+q^a)(1+q^{3b})(1+q^{a+3b}+q^{2a+6b})}$$

and the theorem will be proved if we can show that $H_{a,b} = R_{a,b}$ where $R_{a,b} = R'_{a,b}/W(q^a, q^{3b})$.

A simple calculation gives that

$$R_{a,b} = \frac{(q;q)_{3a+3b}(q;q)_{3b}(q;q)_{2a-1}(q^3;q^3)_{a+3b-1}(q^3;q^3)_{2b-1}(q^3;q^3)_a}{(q;q)_{2a+3b}(q;q)_{a+3b-1}(q;q)_a(q;q)_{a-1}} \cdot (q^3;q^3)_{a+2b}(q^3;q^3)_{a+b}(q^3;q^3)_b(q^3;q^3)_{b-1}. \tag{9.3}$$

A routine calculation gives ($t = q^a$, $s = q^{3b}$)

$$R_{a+1,b}/R_{a,b} = \frac{(1-qt^3s)(1-q^2t^3s)(1-t^2)(1-qt^2)(1-t^3s^3)(1-q^3t^3)}{(1-qt^2s)(1-q^2t^2s)(1-ts)(1-qt)(1-t)(1-q^3t^3s^2)}. \tag{9.4}$$

So let

$$\begin{aligned} F_{a,b}(x) = &\left(\frac{x}{y};q\right)_a \left(\frac{z}{y};q\right)_a \left(\frac{z}{x};q\right)_a \left(\frac{z^2}{xy};q^3\right)_b \left(\frac{xz}{y^2};q^3\right)_b \left(\frac{yz}{x^2};q^3\right)_b \\ &\cdot \left(q\frac{y}{x};q\right)_{a-1} \left(q\frac{y}{z};q\right)_{a-1} \left(q\frac{x}{z};q\right)_{a-1} \\ &\cdot \left(q^3\frac{xy}{z^2};q^3\right)_{b-1} \left(q^3\frac{y^2}{xz}q^3\right)_{b-1} \left(q^3\frac{x^2}{yz};q^3\right)_{b-1}, \end{aligned} \tag{9.5}$$

$$H_{a,b} = \text{C.T. } F_{a,b}.$$

We must show that $H_{a,b} = R_{a,b}$. This will be done by induction on a. First we must prove the base case $H_{1,b} = R_{1,b}$.

Proof of the base case $a = 1$. Substituting $a = 1$ in $R_{a,b}$ given in (9.3) and setting $Q = q^3$ gives

$$R_{1,b} = \frac{(Q)_{3b}}{(Q)_b^3} \frac{(1-Q^b)(1-Q)}{(1-Q^{2b})(1-Q^{2b+1})}. \tag{9.6}$$

We will need the A_2 results proved in § 7. For our present purposes it is convenient to rewrite it in the "fundamental roots" form (sometimes used by Morris [Mo]) obtained by setting $u_1 = x_1/x_2$ and $u_2 = x_2/x_3$. Also let us replace q by Q (so everything is to base Q: $(u_1)_b = (u_1;Q)_b$, etc).

Let $\hat{F}_b$ be defined by

$$\hat{F}_b = (u_1)_b(u_2)_b(u_1u_2)_b(Q/u_1)_{b-1}(Q/u_2)_{b-1}(Q/u_1u_2)_{b-1}. \tag{9.7}$$

Then the results from § 7 that we need here are

$$\text{C.T. } \hat{F}_b = \frac{(Q)_{3b-1}}{(Q)_{b-1}^2 (Q)_b (1-Q^{2b})} \tag{9.8a}$$

and from (7.4) ($A(2, 0, -2)$)

$$\text{C.T. } (u_1 u_2 \hat{F}_b) = \frac{(Q^b - Q)(1+Q^b)}{(1-Q^{2b+1})} (\text{C.T. } \hat{F}_b). \tag{9.8b}$$

Now we want $H_{1,b} = \text{C.T. } F_{1,b}$, where (plug in $a=1$ in (9.5)),

$$F_{1,b} = (1-x/y)(1-z/y)(1-z/x) \cdot \left(\frac{z^2}{xy}; q^3\right)_b \left(\frac{xz}{y^2}; q^3\right)_b \left(\frac{yz}{x^2}; q^3\right)_b \left(q^3 \frac{xy}{z^2}; q^3\right)_{b-1} \left(q^3 \frac{y^2}{xz}; q^3\right)_{b-1} \left(q^3 \frac{x^2}{yz}; q^3\right)_{b-1}.$$

Now let $u_1 = xz/y^2$, $u_2 = yz/x^2$. Then

$$\frac{x}{y} = u_1^{1/3} u_2^{-1/3}, \qquad \frac{z}{y} = u_1^{2/3} u_2^{1/3}, \qquad \frac{z}{x} = u_1^{1/3} u_2^{2/3},$$

and then if we take $Q = q^3$,

$$F_{1,b} = (1 - u_1^{1/3} u_2^{-1/3})(1 - u_1^{2/3} u_2^{1/3})(1 - u_1^{1/3} u_2^{2/3}) \hat{F}_b.$$

So

$$H_{1,b} = \text{C.T. } [1 - u_1^{1/3} u_2^{-1/3} - u_1^{1/3} u_2^{2/3} + u_1 + u_1 u_2 - u_1^{4/3} u_2^{2/3}] \hat{F}_b = \text{C.T. } [(1+u_1 u_2) \hat{F}_b].$$

(u_1 corresponds to the vector $(1, -1, 0) + \delta = (1, -1, 0) + (1, 0, -1) = (2, -1, -1)$, a bad guy (for A_2), and all other terms are even worse: they are fractional. Their contribution is of course zero since $\hat{F}_b$ does not have any terms with fractional exponents, being a Laurent *polynomial.*)

Using (9.8a) and (9.8b) we get

$$H_{1,b} = \left[1 + \frac{(Q^b - Q)(1+Q^b)}{(1-Q^{2b+1})}\right] \frac{(Q)_{3b-1}}{(Q)_{b-1}^2 (Q)_b (1-Q^{2b})}, \tag{9.9}$$

which after a routine calculation turns out to be equal to $R_{1,b}$ in (9.6) (end of proof of the base case $a=1$). □

Proof of the inductive step. Now that we know that $H_{a,b} = R_{a,b}$ for $a=1$ we go next to the inductive step.

For the root system G_2, δ, one-half the sum of the positive roots, is equal to $(-1, -2, 3)$, and (8.6) becomes (recall $t = q^a$)

$$\begin{aligned} H_{a+1,b} &= \text{C.T. } [x^{-1} y^{-2} z^3 G_{a+1,b}] \\ &= \text{C.T. } [x^{-1} y^{-2} z^3 (1-tx/y)(1-tz/y)(1-tz/x) \\ &\qquad \cdot (1-ty/x)(1-ty/z)(1-tx/z) G_{a,b}]. \end{aligned} \tag{9.10}$$

We now expand

$$x^{-1} y^{-2} z^3 (1-tx/y)(1-tz/y)(1-tz/x)(1-ty/x)(1-ty/z)(1-tx/z),$$

discard all the bad guys and collect all the good guys into orbits under W. We then substitute everything back into (9.10) and use the Crucial Lemma, and then finally we

collect terms. (I highly recommend that the reader check this either by hand or by machine, there are only $2^6 = 64$ terms in the expansion.)

We get for (8.7)

$$\begin{aligned} H_{a+1,b} &= (1+t+t^2+t^4+t^5+t^6)\ \text{C.T.}\,[x^{-1}y^{-2}z^3 G_{a,b}] \\ &\quad -(t+t^3+t^5)\ \text{C.T.}\,[x^{-1}y^{-3}z^4 G_{a,b}] \\ &\quad +(t^2+t^3+t^4)\ \text{C.T.}\,[x^{-2}y^{-3}z^5 G_{a,b}] - t^3\ \text{C.T.}\,[x^{-1}y^{-4}z^5 G_{a,b}]. \end{aligned} \tag{9.11}$$

In preparation for the MACSYMA *input file* given below let us put

$$\begin{aligned} x0 &= \text{C.T.}\,[x^{-1}y^{-2}z^3 G_{a,b}]/H_{a,b} \equiv 1 \quad \text{(by definition)}, \\ x1 &= \text{C.T.}\,[x^{-1}y^{-3}z^4 G_{a,b}]/H_{a,b}, \\ x2 &= \text{C.T.}\,[x^{-2}y^{-3}z^5 G_{a,b}]/H_{a,b}, \\ x3 &= \text{C.T.}\,[x^{-1}y^{-4}z^5 G_{a,b}]/H_{a,b}. \end{aligned}$$

With $p0$, $p1$, $p2$, $p3$ as defined in the *input file* below, (9.11) becomes

$$H_{a+1,b}/H_{a,b} = p0*x0 + p1*x1 + p2*x2 + p3*x3 \tag{9.12}$$

and we will call this "sum" in the *input file* below.

Finally we need linear equations relating $x0$, $x1$, $x2$, and $x3$. The simplest vector $z = (z_1, z_2, z_3)$ that makes (8.8) a rational function is $(2, 1, 0)$. Now (8.9) becomes

$$\frac{G_{a,b}(q^2x, qy, z)}{G_{a,b}(x, y, z)} = \frac{P}{Q}$$

where P and Q are computed using (5.3) and (5.4) as modified in § 8 before (8.9). Proceeding as described in § 6 (I used a computer but it is possible to do it by hand) we get the following results. ($a00, \cdots, a23$ are given in the MACSYMA *input file* below.) The choice $\beta = (2, 2, -4)$ yields

$$E_{(2,2,-4)}: \quad a00*x0 + a01*x1 = 0;$$

$\beta = (0, 3, -3)$ yields

$$E_{(0,3,-3)}: \quad a10*x0 + a11*x1 + a12*x2 = 0;$$

and $\beta = (4, 0, -4)$ yields

$$E_{(4,0,-4)}: \quad a20*x0 + a21*x1 + a22*x2 + a23*x3 = 0.$$

(A copy of the *C* program that implements the algorithm of § 5, modified as in § 8, by which I obtained the above equations, is available upon request (either a printout by U.S. mail or by electronic mail; sorry, no disks). However it is highly recommended that the readers write their own programs. It is much easier to write your own code than to try to understand somebody else's computer scratch.)

MACSYMA INPUT FILE

```
a00: t^3*s^2*q - t/q + t*s - t^3*s + t^2*s^2 - t^4*s*q - t^2 + s/q
+ t^2*s*q - t^2*s/q + 1 - t^4*s^2$
a00: 0 - a00$
a01: t^4*s^2*q - q^ - 1$
a10: -t^4*s^2*q + q^2 + t^2*s^2*q + t^2*s^2 - t^4*s*q
```

```
        -t^2*q^3-t^2*q^2+s*q^2+s-t^2-t^4*s*q^3+t^2*s^2*q^3
        +t*s^2+t^2*s*q-t^3*s*q-t^3*s-t^3*q^3-t^2*s*q^2
        +t*s*q^3+t*s*q^2+2*t*s-t^3*q-t^3-2*t^3*s*q^3
        +t*s^2*q^3+t*s^2*q^2$
a11: t^2*s*q+t^2*s-t^2*s*q^3-t^2*s*q^2-s+t^2*q+t^2
        +t^4*s*q^3-t^2*s^2*q^3-t^2*s^2*q^2+t-t^3*s^2*q^3$
a12: t-t^3*s^2*q^3+1-t^4*s^2*q^3$
a12: (-1)*a12$
a20: t*s^2*q-t^2*s^2-t^3*q^-3+t^2*q^-2-t^2*s*q+t^3*s+t^2*s*q^-3
        -t*s*q^-2-t*s+t^3*s*q^-2$
a21: -t^2*s^2*q+t^3*s^2*q+t^3*s^2+t^2*q^-3-t*q^-2-t*q^-3
        +t^3*s*q-t*s*q^-3$
a22: t^3*s^2*q-t*q^-3$
a23: -t^4*s^2*q+q^-3$
p0: 1+t+t^2+t^4+t^5+t^6$
p1: -t-t^3-t^5$
p2: t^2+t^3+t^4$
p3: -t^3$
x0: 1$
x1: 0-a00*x0$
x1: x1/a01$
x2: a10*x0+a11*x1$
x2: 0-x2/a12$
x3: a20*x0+a21*x1+a22*x2$
x3: 0-x3/a23$
sum: p0*x0+p1*x1+p2*x2+p3*x3$
rhs: (1-q*t^3*s)*(1-q^2*t^3*s)*(1-t^2)*(1-q*t^2)$
rhs: rhs*(1-t^3*s^3)*(1-q^3*t^3)$
rhs: rhs/((1-q*t^2*s)*(1-q^2*t^2*s)*(1-t*s)*(1-q*t)*(1-t)*
        (1-q^3*t^3*s^2))$
sum: sum-rhs$
ratsimp(sum);
quit( );
```

In the *input file* we solve for $x1, x2, x3$ successively. Then we ask MACSYMA to compute "sum" $= H_{a+1,b}/H_{a,b}$. We enter $R_{a+1,b}/R_{a,b}$ given in (9.4) and call it "rhs." So far every line has been terminated with a dollar sign so that the partial steps are not going to be printed out. The second line from the bottom is

$$sum: sum - rhs\$$$

that defines the new "sum" to be the difference between the conjectured right-hand side and the real right-hand side. This should be zero if the conjecture is true. The last line asks MACSYMA to simplify this difference: *ratsimp*(*sum*); and now, finally, there is a semicolon, because now we *do* want to see the answer.

On December 22, 1986, 3:30 p.m., after two previous unsuccessful attempts (due to typing errors that were presently detected), I typed on my terminal:

$$macsyma < inputfile$$

After a few minutes came the output: 27 blank double lines (due to the dollar signs) and

$(c28)$
$(d28)$ 0

YEA!!!

□

10. Prospects. The next in line is F_4. But before we can do F_4, we must do D_4, the short part of F_4. A preliminary calculation done by Dave Robbins shows that for D_4, (4.6) involves more than one hundred terms. So we will have to find and solve a system of more than one hundred linear equations with rather complicated coefficients. While this is still within the reach of current computers, it is hard to justify that kind of expense before all other means have been exhausted.

As I have already mentioned, the reason why the Macdonald-Morris conjectures (qM-M2) are easier than the original Macdonald conjectures (qM), is that the two parameters let us break the problem into two subproblems. In a way we are first doing the long roots and only then the short roots. But nowhere in § 8 have we ever used the "physical appearance" of the short and long roots, that is the fact that the roots of R_{long} are "longer" than those of R_{short}. All we used was the fact that the partition $R = R_{\text{short}} \cup R_{\text{long}}$ partitions the root-system R into two subsets both of which are invariant under the action of the Weyl group W.

Is it possible to find such a partition for those root systems that have only one root length (A_n, D_n, E_6, E_7, E_8)? The answer is: not quite, but almost. Instead of the Weyl group W itself, we have to settle for invariance under a certain subgroup of W. It turns out that it is possible to find such a partition of R which is invariant under a very large subgroup of W, so we only have to sacrifice a little bit of symmetry. Still, we have to put up with some vectors that were previously denounced as bad guys. In return, however, the corresponding polynomial that appears in (4.1) is much smaller and the trade-off is well in our favor, since the resulting set S in (4.6) turns out to be much smaller.

For example, A_{n-1} can be partitioned into

$$A_{n-1} = \{\pm(e_1 - e_i);\ 2 \leqq i \leqq n\} \cup \{\pm(e_i - e_j);\ 2 \leqq i < j \leqq n\}.$$

The second set is the subroot system A_{n-2} in the last $n-1$ coordinates and its Weyl group S_{n-1} (that acts by permuting the last $n-1$ coordinates) is the subgroup that leaves both subsets invariant.

In fact, the first subset above can be further partitoned into its positive and negative roots and so A_{n-1} can be partitioned into *three* subsets, each of which is invariant under the above-mentioned S_{n-1}. Indeed, we have

$$A_{n-1} = \{e_1 - e_i;\ 2 \leqq i \leqq n\} \cup \{-e_1 + e_i;\ 2 \leqq i \leqq n\} \cup \{\pm(e_i - e_j;\ 2 \leqq i < j \leqq n)\}.$$

We should thus expect a three parameter "pseudo"-Macdonald-Morris conjecture

$$\text{(10.1)}\quad \text{C.T.} \prod_{i=2}^{n} (x_1/x_i)_a \prod_{i=2}^{n} (qx_i/x_1)_b \prod_{2 \leqq i < j \leqq n} (x_i/x_j)_c (qx_j/x_i)_{c-1} = \text{something explicit.}$$

Such an identity indeed exists and was conjectured by Morris [Mo] (Morris proved the $q=1$ case). It was recently proved by Kadell [Kad1] and Habsieger [Hab2], who deduced it from their Askey q-Selberg integral. However it is possible to get a Stembridge-style proof by using the method of § 8 [Z4]. The $a=0$ case is easily seen to be equivalent to the $a=b=0$ case. Then one inducts on a and gets a recurrence in a. The analogue of (4.6) contains only n terms and it is easy to find $n-1$ independent equations satisfied by them. The base case $a=b=0$ is just the A_{n-1} case while the $a=b=c$ is the A_n case. This provides the necessary induction.

For D_n the situation is not quite as rosy, but it is still very promising. While it is not possible to partition D_n into *three* subsets invariant under a large subgroup of W, it is possible to do it with *two* subsets.

Indeed,

$$D_n = \{\pm e_1 \pm e_i;\ 2 \leqq i \leqq n\} \cup \{\pm e_i + e_j;\ 2 \leqq i < j \leqq n\}. \tag{10.2}$$

The second set is just the root-system D_{n-1} on the last $n-1$ coordinates. The Weyl group of this D_{n-1} consists of all signed permutations with an even number of signs that act on the last $n-1$ coordinates [Bo, p. 257, (X)]. It leaves both subsets of (10.2) invariant. I conjecture that if

$$\begin{aligned} H_{a,b} =: \text{C.T.} \prod_{i=2}^{n} (x_1/x_i)_a (qx_i/x_1)_{a-1} (x_1 x_i)_a (q/x_1 x_i)_{a-1} \\ \cdot \prod_{2 \leqq i < j \leqq n} (x_i/x_j)_b (qx_j/x_i)_{b-1} (x_i x_j)_b (q/x_i x_j)_{b-1}, \end{aligned} \tag{10.3}$$

then $H_{a,b}$ has an explicit and perhaps nice expression. In any case the method described in § 8 should produce $H_{a,b+1}/H_{a,b}$, a certain rational function, and whether it is nice or not it should give us a formula for $H_{a,b}$ that we know should be nice when $a = b$. In any case the analogue of (4.6) is much simpler now, and the number of equations needed is considerably reduced. The base case $a = 1$ is essentially D_{n-1}, and once we obtain the recurrence in a, and thus the expression for $H_{a,b}$, then $H_{b,b}$ will give the D_n case of (qM). Once we will have D_n, the remaining classical families B_n and C_n should be relatively easy. D_n is the "hard core" of both B_n and C_n, and it hopefully would be relatively easy to add the rest. Similarly, it should be possible to find more refined conjectures for F_4 and the E's that will enable us to break the proof into manageable parts.

Another possibility is to find the "trivializing generalization": a much more general statement than the Macdonald conjectures that would be trivial (or at least easy, or in any case possible) to prove. Except for some coefficients in the A_{n-1} case [Ste3], the general coefficients of G_k and $G_{a,b}$ do not seem to have nice expressions. So we have to abandon the hope of finding a *nice* expression for the general coefficient of G_k. But perhaps it is possible to find certain *linear combinations* of these messy coefficients that are good-looking. Remember that in our method the desired coefficient, H_{k+1}, was obtained, via (4.6) as a certain linear combination of more or less ugly coefficients of the k case. Maybe it is possible to find a family of polynomials, a_λ, say, parametrized by partitions λ such that

$$\text{C.T.} \left[\prod_{\alpha \in R^+} (x^\alpha)_k (qx^{-\alpha})_{k-1} a_\lambda \right] \tag{10.4}$$

has a nice expression in k and λ. Now that we have a laboratory for producing not only the constant term, but also other coefficients of $F_k(F_{a,b})$, there is a vast hunting ground for formulating and testing such more general conjectures (see [Kad2] for a similar idea in the context of the Selberg integral).

A related idea, inspired by Aomoto's [Ao] proof of Selberg's integral, was suggested by Askey [As3]: Break the ascent from k to $k+1$ in (qM) by raising the subscripts on the roots one, or few, at a time. The present method also offers a convenient workbench for Askey's approach. In particular it is possible to verify his G_2 conjectures made at the end of § 4 of [As3].

Acknowledgments. I wish to thank all the people, machines, and institutions that helped me in this research. Among the first category I wish to thank: John Stembridge

and Dennis Stanton for the brilliant ideas that led to this paper; Richard Askey, Dominique Foata, and Dave Robbins for many stimulating discussions; Chip Morris for his extraordinary thesis [Mo] that enabled a mere mortal to understand the Macdonald conjectures; Dave Robbins (again) for independently verifying, using ALTRAN, the MACSYMA calculations in § 9; and, of course, I. G. Macdonald for his intriguing conjectures that have been keeping me busy all these years. I also wish to thank my wife, Jane, for convincing me that C is a better language than BASIC, and Marci Perlstadt and Ron Perline for initiating me to MACSYMA. Finally, the referee deserves thanks for many helpful remarks.

In the second category, I wish to thank the Drexel Electrical Engineering Department's VAX 780 for being such a faithful slave and for its hospitality toward the time-consuming MACSYMA.

Among the third category I wish to thank the Drexel Electrical Engineering Department for letting a poor cousin from the Mathematics Department use their computer.

Note added in proof. Kevin Kadell has meanwhile proved the BC_n cases of (qM-M1), and thus also the B_n, C_n, D_n cases. Frank Garvan (preprint) has used the method of this paper to prove the $q=1$ case of F_4. He also succeeded in proving the F_4, I_3 cases of the Macdonald-Mehta conjectures. Frank Garvan and Dennis Stanton have proved that the system (5.9) is always upper triangular, in the $q=1$ case.

REFERENCES

[A1] W. ALLEN, *A giant step for mankind*, in Side Effects, Ballantine Books, New York, 1981, pp. 127-138.

[An1] G. ANDREWS, *Problems and prospects for basic hypergeometric functions*, in The Theory and Applications of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 191-224.

[An2] ———, *q-Series: Their Development and Applications in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra*, CBMS Regional Conference Series in Mathematics 66, American Mathematical Society, Providence, RI, 1986.

[Ao] K. AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.

[As1] R. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.

[As2] ———, *A q-beta integral associated with* BC_1, SIAM J. Math. Anal., 13 (1982), pp. 1008-1010.

[As3] ———, *Integration and Computers*, Proceedings of a Computer Algebra Conference, D. Chudnovsky and G. Chudnovsky, eds., to appear.

[As4] ———, *Séance de problèmes*, in Combinatoire Enumerative, G. Labelle and P. Leroux, eds., Lecture Notes in Mathematics 1234, Springer-Verlag, Berlin, 1986, pp. 381-382.

[Bo] N. BOURBAKI, *Groupes et algèbres de Lie, chapitres* IV, V, VI, Herman, Paris, 1968.

[Br1] D. BRESSOUD, *A Colored tournaments and Weyl's denominator formula*, Pennsylvania State University, preprint.

[Br2] ———, *Definite integral evaluation by enumeration*, in Combinatoire Enumerative, G. Labelle and P. Leroux, eds., Lecture Notes in Mathematics 1234, Springer-Verlag, Berlin, 1986, pp. 48-57.

[B-G] D. BRESSOUD AND I. GOULDEN, *Constant term identities extending the q-Dyson theorem*, Trans. Amer. Math. Soc., 291 (1985), pp. 203-228.

[C-H] R. CALDERBANK AND P. HANLON, *An extension to root-systems of a theorem on tournaments*, Combin. Theory Ser. A, 41 (1986), pp. 228-245.

[C] R. W. CARTER, *Simple Groups of Lie Type*, John Wiley, London, New York, 1972.

[C-F] P. CARTIER AND D. FOATA, *Problèmes combinatoires de commutation et réarrangements*, Lecture Notes in Mathematics 85, Springer-Verlag, Berlin, 1969.

[C-R] P. COHEN AND A. REGEV, *Asymptotics of combinatorial sums and the central limit theorem*, SIAM J. Math. Anal., 19 (1988), to appear.

[D1] F. J. DYSON, *Statistical theory of the energy levels of complex systems* I., J. Math. Phys., 3 (1962), pp. 140-156.

[D2] ———, *Missed opportunities*, Bull. Amer. Math. Soc., 78 (1972), pp. 635-653.

[E] R. EVANS, *Character sum analogues of constant term identities for root systems*, Israel J. Math, 46 (1983), pp. 189-196.

[F] D. FOATA, *Etude algébrique de certains problèmes d'analyse combinatoire et du calcul des probabilités*, Publ. Inst. Statist. Univ. Paris., 14 (1965), pp. 81-241.

[Ge] I. GESSEL, *Tournaments and Vandermonde's determinant*, J. Graph Theory, 3 (1979), pp. 305-307.

[Goo] I. J. GOOD, *Short proof of a conjecture of Dyson*, J. Math. Phys., 11 (1970), p. 1884.

[Gou] I. GOULDEN, *Macdonald's constant term for A_{n-1} and the inner and internal product for symmetric functions*, Report 86-22, Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada, 1986.

[Gr] J. R. GREENE, *Bijections for permutation statistics*, Discrete Math., to appear.

[Gu] J. GUNSON, *Proof of a conjecture of Dyson in the statistical theory of energy levels*, J. Math. Phys., 3 (1962), pp. 752-753.

[Hab1] L. HABSIEGER, *La q-conjecture de Macdonald-Morris pour G_2*, C.R. Acad. Sci., 303 (1986), pp. 211-213.

[Hab2] ———, *Une q-intégrale de Selberg-Askey*, SIAM J. Math. Anal., 19 (1988), to appear.

[Han1] P. HANLON, *The proof of a limiting case of Macdonald's root system conjectures*, Proc. Lond. Math. Soc., 49 (1984), pp. 170-182.

[Han2] ———, *On the decomposition of the tensor algebra of the classical Lie algebras*, Adv. Math., 56 (1985), pp. 238-282.

[Han3] ———, *Cyclic homology and the Macdonald conjectures*, Inv. Math., 86 (1986), pp. 131-159.

[Kac] G. KAC, *Infinite dimensional Lie algebras*, 2nd ed., Cambridge University Press, Cambridge, 1985.

[Kad1] K. KADELL, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969-986.

[Kad2] ———, *The q-Selberg polynomials for $n = 2$*, preprint.

[Ma1] I. G. MACDONALD, *Affine root systems and Dedekind's η-function*, Inv. Math., 15 (1972), pp. 91-143.

[Ma2] ———, *The Poincaré series of a Coxeter group*, Math. Anal., 199 (1972), pp. 161-174.

[Ma3] ———, *Some conjectures for root systems*, SIAM J. Math. Anal., 13 (1982), pp. 988-1007.

[Me] M. L. MEHTA, *Random Matrices and Statistical Theory of Energy Levels*, Academic Press, New York, 1967.

[Mi] S. C. MILNE, *An elementary proof of the Macdonald identities for $A_l^{(1)}$*, Adv. Math., 57 (1985), pp. 34-70.

[Mo] W. G. MORRIS II, *Constant term identities for finite and affine root systems*, Ph.D. thesis, Univ. of Wisconsin, Madison, WI (available from University Microfilm, Ann Arbor, MI).

[R1] A. REGEV, *Asymptotic values for degrees associated with strips of Young diagrams*, Adv. Math., 41 (1981), pp. 115-136.

[R2] ———, *Combinatorial sums, identities and trace identities of the 2×2 matrices*, Adv. Math., 46 (1982), pp. 230-240.

[Se] A. SELBERG, *Bemerkninger om et multiplet integral*, Norske Mat. Tidsskr., 26 (1944), pp. 71-78.

[Stanl1] R. STANLEY, *The q-Dyson conjecture, generalized exponents, and the internal product of Schur functions*, in Combinatorics and Algebra, C. Greene, ed., Contemporary Mathematics 34, American Mathematical Society, Providence, RI, 1984, pp. 81-94.

[Stanl2] ———, *The stable behavior of some characters of $SL(n, C)$*, Linear and Multinear Algebra., 16 (1984), pp. 3-27.

[Stant1] D. STANTON, *Sign variations of the Macdonald identities*, Report 84-130, University of Minnesota, Minneapolis, MN, 1984.

[Stant2] ———, *Sign variations of the Macdonald identities*, SIAM J. Math. Anal., 17 (1986), pp. 1454-1460.

[Ste1] J. STEMBRIDGE, *Combinatorial Decompositions of Characters of $SL(n, C)$*, thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1985 (available from University Microfilm, Ann Arbor, MI).

[Ste2] ———, *First layer formulas for the characters of $SL(n, C)$*, Trans. Amer. Math. Soc., to appear.

[Ste3] ———, *A short proof of Macdonald's conjecture for the root systems of type A*, Proc. Amer. Math. Soc., to appear.

[W] K. WILSON, *Proof of a conjecture of Dyson*, J. Math. Phys., 3 (1962), pp. 1040-1043.

[Z1] D. ZEILBERGER, *A combinatorial proof of Dyson's conjecture*, Discrete Math., 41 (1982), pp. 317-321.

[Z2] ———, *A proof of the G_2 case of Macdonald's root system-Dyson conjecture*, SIAM J. Math. Anal., 18 (1987), pp. 880-883.

[Z3] ———, private communication.

[Z4] ———, *A Stembridge-Stanton style elementary proof of the Habsieger-Kadell q-Morris identity*, Discrete Math., to appear.

[Z-B] D. ZEILBERGER AND D. BRESSOUD, *A proof of Andrews' q-Dyson conjecture*, Discrete Math., 54 (1985), pp. 201-224.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

STRONG INTERNAL RESONANCE, $Z_2 \oplus Z_2$ SYMMETRY, AND MULTIPLE PERIODIC SOLUTIONS*

THOMAS J. BRIDGES†

Abstract. Equal (strong internal resonance) or nearly equal natural frequencies in a Hamiltonian system are shown to lead to a higher multiplicity and secondary branching of periodic solutions. The role of $Z_2 \oplus Z_2$ symmetry is emphasized as a basis for the analysis. The Lyapunov-Schmidt method is used to generate a set of bifurcation equations. The higher order terms in the bifurcation equations are formally neglected, leading to the basic normal form for coupled equations with $Z_2 \oplus Z_2$ symmetry. Known results for this normal form are used to determine the nature of the periodic solutions in a neighborhood of a 1 : 1 resonance. A stability analysis is based on Floquet theory and the Lyapunov-Schmidt method. Regions of stability are established and particular singular points in parameter space are found where periodic solutions may not exist. The analysis is carried out on an example: the orthogonal planar pendulum.

Key words. Hamiltonian system, bifurcation, symmetry, singularity theory

AMS(MOS) subject classifications. 34C15, 34C25, 47H15, 70K99

1. Introduction. The circle group S^1 and the symmetry group Z_2 play a fundamental role in an analysis of the nature of periodic solutions of a Hamiltonian system near equilibrium. The action of the circle group represents the translation invariance in time. A development of this role with the use of the Lyapunov-Schmidt method has been given by Moser [23]. The role of Z_2 symmetry is more subtle and appears in the bifurcation equations *after* application of a Lyapunov-Schmidt splitting. The importance of Z_2 symmetry and its basic role in the analysis of periodic solutions has been emphasized by Golubitsky and Langford [10]. In Hamiltonian systems, where a Hopf bifurcation parameter is absent, it is usual to treat the period as a bifurcation parameter and then apply the Lyapunov-Schmidt method (Hale [13, Chap. 8]). For a nonresonant Hamiltonian in this setting it is generic that the bifurcation equations are Z_2 equivalent to the pitchfork. When a 1 : 1 resonance occurs we find that the symmetry group $Z_2 \oplus Z_2$ plays a central role. Similarly the symmetry group $Z_2 \oplus Z_2 \oplus Z_2$ can be shown to play a central role in an analysis of the $1:1:k$ $(k \geqq 1)$ resonance.

This is similar to the role played by $Z_2 \oplus Z_2$ symmetry when a double eigenvalue occurs in an equilibrium problem. In this context Bauer, Keller, and Reiss [1] first showed that splitting a double eigenvalue results in secondary bifurcation. Golubitsky and Schaeffer [11] subsequently showed that this was a result of the unfolding of the normal form for equations with $Z_2 \oplus Z_2$ symmetry. This principle of splitting a double eigenvalue to find secondary branches has since been observed in a number of interesting applications. Golubitsky and Schaeffer [12] have analyzed the buckling of rectangular plates, Buzano [5] has analyzed the buckling of thin rods, and Kriegsmann and Reiss [17] have analyzed magneto-hydrodynamic equilibria. Margolis and Matkowsky [20] first applied this concept to the secondary branching of periodic solutions, and Bridges [3], [4] subsequently applied this theory to find secondary branches of periodic surface waves in an enclosed basin.

From a strictly Hamiltonian point of view, periodic solutions near a resonant equilibrium have been well studied in the celestial mechanics literature. Hénon and Heiles [14], Braun [2], and Kummer [18], for example, have used numerical methods and the Gustavson normal form to determine the multiplicity and stability of periodic

* Received by the editors, October 29, 1986; accepted for publication August 24, 1987.
† Department of Mathematics, Worcester Polytechnic Institute, Worcester, Massachusetts 01609.

solutions *at* a 1:1 resonance. An up-to-date review of generic bifurcations in Hamiltonian systems is given by Meyer [21].

In this paper the Lyapunov-Schmidt method is used to reduce the analysis of Hamiltonian systems at 1:1 resonance to a set of bifurcation equations. Known properties of the normal form for $Z_2 \oplus Z_2$ symmetry are then used to analyze the nature and multiplicity of the periodic solutions. Stability is determined using Floquet theory. The number of degrees of freedom may be arbitrary as long as the other degrees of freedom are nonresonant. Clearly it is sufficient to consider a system with two degrees of freedom,

$$H(\theta_1, \dot\theta_1, \theta_2, \dot\theta_2) = \tfrac{1}{2}\omega_1(\theta_1^2 + \dot\theta_1^2) + \tfrac{1}{2}\omega_2(\theta_2^2 + \dot\theta_2^2) + H_3$$

where H_3 is a convergent power series that begins with a third-order term.

However, for added insight, we will develop the ideas by considering a physical problem with two degrees of freedom, which is sufficiently general and contains some interesting properties besides. The system is a compound orthogonal pendulum.

Consider a right-handed coordinate system with the z-axis directed upwards and the x-axis directed to the right with a pendulum of mass m_1 and length l_1 suspended from the origin with its motion restricted to the x-z-plane. Suspended from the mass m_1 is a second pendulum of mass m_2 and length l_2 *whose motion is restricted to the y-z-plane.* In addition the base of the pendulum (the origin) undergoes a horizontal planar harmonic excitation, say

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} \mu_1 l_1 2\sqrt{2} \cos(\omega t + \zeta) \\ \mu_2 l_2 2\sqrt{2} \cos \omega t \end{pmatrix} \tag{1.1}$$

where $\mu = (\mu_1, \mu_2)$ is a measure of the excitation amplitude and ζ corresponds to the phase between the two components of the excitation.

The Lagrangian for such a system is easily shown to be

$$\begin{aligned} \mathbf{L} = {} & \tfrac{1}{2}(m_1 + m_2)[\dot x_0^2 + \dot y_0^2 + 2 l_1 \dot x_0 \dot\theta_1] + m_2 l_2 \dot y_2 \dot\theta_2 \cos\theta_1 \cos\theta_2 \\ & + \tfrac{1}{2}(m_1 + m_2) l_1^2 \dot\theta_1^2 + \tfrac{1}{2} m_2 l_2^2 \dot\theta_2^2 + m_2 l_1 l_2 \dot\theta_1 \dot\theta_2 \sin\theta_1 \sin\theta_2 \\ & + (m_1 + m_2) g l_1 (\cos\theta_1 - 1) + m_2 g l_2 (\cos\theta_2 - 1). \end{aligned} \tag{1.2}$$

The interval of time $[0, 2\pi/\omega]$ is mapped to $[0, 2\pi]$ and we define the parameters $\lambda = g/\omega^2 l_1$, $l_1/l_2 = 1 - \sigma$ and $m = m_2/(m_1 + m_2)$. Then application of the Euler-Lagrange operator to (1.2) generates the nonlinearly coupled pair of second-order ordinary differential equations

$$\begin{aligned} & \frac{d^2\theta_1}{dt^2} + \lambda \sin\theta_1 - \mu_1 2\sqrt{2} \cos(t + \zeta) \cos\theta_1 \\ & \qquad + \frac{m}{1 - \sigma} \sin\theta_1 [\ddot\theta_2 \sin\theta_2 + \dot\theta_2^2 \cos\theta_2] = 0, \end{aligned} \tag{1.3a}$$

$$\begin{aligned} & \frac{d^2\theta_2}{dt^2} + \lambda(1 - \sigma) \sin\theta_2 - \mu_2 2\sqrt{2} \cos t \cos\theta_2 \\ & \qquad + (1 - \sigma) \sin\theta_2 [\ddot\theta_1 \sin\theta_1 + \dot\theta_1^2 \cos\theta_1] = 0 \end{aligned} \tag{1.3b}$$

where λ will play the role of a bifurcation parameter, σ is a parameter that is used to unfold the 1:1 resonance, and $\mu \in R^2$ will be an unfolding parameter that breaks the symmetry. We are interested in periodic solutions of the set of equations (1.3) when the two linear natural frequencies are equal or nearly equal, that is, in the neighborhood of $\sigma = 0$. Physically this is when the lengths of the two pendulums are very nearly equal.

Although the results herein are for the 1:1 resonance they are not so limited. The $1:k$ resonance, although referred to as a higher order resonance when $k \geqq 1$, behaves much like a 1:1 resonance. It can be shown that the $1:k$ resonance may have a normal form similar to that of the 1:1 resonance and in addition the $1:1:k$ resonance can be analyzed using the normal form for a $Z_2 \oplus Z_2 \oplus Z_2$ symmetric problem.

The results herein for secondary branching of periodic solutions may also hold for the 1:1 resonance in nondegenerate Hopf bifurcation when two parameters are present. The analogous situation is given by the following evolution equation with $u \in R^4$:

$$\omega \frac{du}{dt} + \mathbf{A}(\sigma)u + \mathbf{B}(h)u = \mathbf{F}(h, \sigma, u)$$

where $h, \sigma \in R$, $\mathbf{F}(h, \sigma, 0) = 0$ and $\mathbf{F}$ is nonlinear in u, $\mathbf{A}, \mathbf{B} \in R^{4\times 4}$ with $\mathbf{B}(0) = 0$ and $\mathbf{A}(\sigma)$ has eigenvalues $\lambda_1^-(\sigma), \lambda_1^+(\sigma), \lambda_2^-(\sigma), \lambda_2^+(\sigma)$ with $\lambda_1^-(0) = -i$, $\lambda_1^+(0) = i$, $\lambda_2^-(0) = -i$, and $\lambda_2^+(0) = i$. If varying σ away from zero splits the 1:1 resonance with h fixed, then we expect analogous secondary branching of periodic or quasi-periodic solutions. Related results for the Hopf bifurcation theorem at resonance have been obtained by Kielhöfer [15], Caprino, Maffei, and Negrini [6], and Chow, Mallet-Paret, and Yorke [7].

2. Symmetry and the branching equation. It is usual to convert (1.3) to a set of four first-order equations and set the problem in a Banach space with a graph norm. With a second-order derivative in the nonlinearity, however, some simplification is found by working from an integral equation point of view. Consider the integral operators

$$\mathbf{K}\phi = \int_0^{2\pi} k(t, \tau)\phi(\tau)\, d\tau, \tag{2.1a}$$

$$\mathbf{K}_t\phi = \int_0^{2\pi} k_t(t, \tau)\phi(\tau)\, d\tau \tag{2.1b}$$

where

$$k(t, \tau) = \sum_{n=1}^{\infty} \frac{\psi_n(t)\psi_n^*(\tau) + \psi_n^*(t)\psi_n(\tau)}{n^2} \tag{2.1c}$$

and $\psi_n(t) = e^{int}/\sqrt{2\pi}$, $i = \sqrt{-1}$; the * denotes complex conjugation and $k_t(t, \tau) = \partial k/\partial t$. Expressed using the above operators, terms in the equations (1.3) become $\ddot{\theta}_j = \phi_j$, $\theta_j = -\mathbf{K}\phi_j$ and $\dot{\theta}_j = -\mathbf{K}_t\phi_j$ for $j = 1, 2$.

The governing equations may then be written as the abstract operation

$$\Phi(\phi, \lambda, \sigma, \mu) = 0$$

for fixed m with $\Phi = (\Phi_1, \Phi_2)$ and $\phi = (\phi_1, \phi_2)$ and the operators Φ_j are given explicitly by

$$\begin{aligned}\Phi_1(\phi, \lambda, \sigma, \mu) &= \phi_1 - \lambda \sin \mathbf{K}\phi_1 - 2\sqrt{2}\mu_1 \cos(t+\zeta) \cos \mathbf{K}\phi_1 \\ &\quad + \frac{m}{1-\sigma} \sin \mathbf{K}\phi_1 [\phi_2 \sin \mathbf{K}\phi_2 - (\mathbf{K}_t\phi_2)^2 \cos \mathbf{K}\phi_2],\end{aligned} \tag{2.2a}$$

$$\begin{aligned}\Phi_2(\phi, \lambda, \sigma, \mu) &= \phi_2 - \lambda(1-\sigma) \sin \mathbf{K}\phi_2 - 2\sqrt{2}\mu_2 \cos t \cos \mathbf{K}\phi_2 \\ &\quad + (1-\sigma) \sin \mathbf{K}\phi_2 [\phi_1 \sin \mathbf{K}\phi_1 - (\mathbf{K}_t\phi_1)^2 \cos \mathbf{K}\phi_1].\end{aligned} \tag{2.2b}$$

Φ is now a compact mapping $\Phi: \chi \times R \times R \times R^2 \to \chi$ where χ is a Banach space of pairs of continuous functions, $\chi = \chi_0 \times \chi_0$ where χ_0 is the Banach space $\chi_0 = \{f: f(t+2\pi) = f(t), f: \mathbf{R} \to \mathbf{C}$ is continuous, *and* $\int_0^{2\pi} f(\tau)\, d\tau = 0\}$ and for any $f \in \chi_0$ the $\|f\| = \sup_t |f(t)|$ and for any $\phi \in \chi$ we have $\|\phi\|_\chi = \max\{\|\phi_1\|, \|\phi_2\|\}$.

When $\mu = 0$ the mapping Φ satisfies two sets of symmetry conditions. Due to the translation invariance in time Φ_1 and Φ_2 commute with the action of the circle group $\Gamma_\theta = S^1$. Associating angles θ in S^1 with numbers in $[0, 2\pi)$,

$$\Gamma_\theta f(t) = f(t-\theta)$$

for $\theta \in \Gamma_\theta$, $f \in \chi_0$. On χ we consider the diagonal action of Γ_θ on $\chi_0 \times \chi_0$. In addition we note that (2.2a) is odd in ϕ_1 and even in ϕ_2 whereas (2.2b) is even in ϕ_1 and odd in ϕ_2. In other words Φ commutes with the action of the symmetry group $\Gamma_2 = Z_2 \oplus Z_2$ [12, Chap. X], where Γ_2 is a four element group that may be represented by the set of four diagonal operators $\Gamma_2 = \text{diag}[\pm 1, \pm 1]$. In summary we have Proposition 2.1.

PROPOSITION 2.1. *When $\mu = 0$, Φ commutes with the action of Γ_θ and Γ_2.*

Proof. The proof comes from the fact that $\Gamma_2 \Phi(\phi, \lambda, \sigma, 0) = \Phi(\Gamma_2 \phi, \lambda, \sigma, 0)$ which follows from substitution of $\Gamma_2 = \text{diag}[\pm 1, \pm 1]$ and noting that Φ_1 is odd in ϕ_1 and even in ϕ_2 and Φ_2 is even in ϕ_1 and odd in ϕ_2. The fact that $\Phi(\phi, \lambda, \sigma, 0)$ commutes with the diagonal action of Γ_θ on $\chi_0 \times \chi_0$ follows if Γ_θ commutes with $\mathbf{K}$. Consider

$$\begin{aligned}
\Gamma_\theta \mathbf{K} \phi_j &= \int_0^{2\pi} k(t-\theta, \tau) \phi_j(\tau)\, d\tau \\
&= \int_0^{2\pi} k(t, \tau+\theta)\, \phi_j(\tau)\, d\tau \\
&= \int_\theta^{\theta+2\pi} k(t, s) \phi_j(s-\theta)\, ds \\
&= \mathbf{K} \Gamma_\theta \phi_j. \qquad \square
\end{aligned}$$

Whether or not Φ commutes with the action of Γ_2 is not a necessary component of the analysis. The important point is that the bifurcation equations commute with action of Γ_2. An example from celestial mechanics is the well-studied Hamiltonian at 1:1 resonance of Hénon and Heiles [14], [2], [18], [19],

$$H = \tfrac{1}{2}(\theta_1^2 + \dot\theta_1^2) + \tfrac{1}{2}(\theta_2 + \dot\theta_2^2) + c\theta_1^2\theta_2 + d\theta_2^3 \tag{2.3}$$

that corresponds to the differential equations

$$\frac{d^2\theta_1}{dt^2} + \theta_1 + 2c\theta_1\theta_2 = 0, \tag{2.4a}$$

$$\frac{d^2\theta_2}{dt^2} + \theta_2 + c\theta_1^2 + 3\, d\theta_2^2 = 0. \tag{2.4b}$$

This pair considered as a mapping does not commute with the action of Γ_2. However the bifurcation equations do. Additional properties of this equation are considered in § 6.

When $\mu \neq 0$ both the Γ_2 and Γ_θ symmetry properties are lost. However for a more general analysis μ will be included in the Lyapunov-Schmidt reduction and when appropriate the symmetry properties are reintroduced in the analysis of the bifurcation equations for the special case $\mu = 0$. For the Lyapunov-Schmidt reduction we need Proposition 2.2.

PROPOSITION 2.2. *$\Phi(\phi, \lambda, \sigma, \mu)$ is continuously Fréchet differentiable with respect to ϕ and the Fréchet derivative of Φ at $\phi = 0$, $\lambda = 1$, and $\sigma = \mu = 0$ is a symmetric Fredholm operator with index zero and four-dimensional nullspace.*

Proof. First define an identity operator on χ by $\mathbf{I} = \operatorname{diag}[1, 1]$. Then the Fréchet derivative Φ_ϕ of Φ at $\phi = 0$, $\lambda = 1$, $\sigma = 0$ acting on a vector $u = (u_1, u_2)$ is

$$\Phi_\phi(0, 1, 0, 0)\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = [\mathbf{I} - \Lambda\mathbf{K}]\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

where in this case Λ is a diagonal operator $\Lambda = \operatorname{diag}[1, 1]$ and $\mathbf{K}$ is as defined in (2.1a). $\mathbf{K}$ is a completely continuous operator with symmetric kernel $k(t, \tau) = k(\tau, t)$. Therefore the dim $N(\mathbf{I} - \Lambda\mathbf{K}) = \operatorname{codim} R(\mathbf{I} - \Lambda\mathbf{K})$. The operator $[\mathbf{I} - \Lambda\mathbf{K}]$ is a diagonal operator acting on $\chi_0 \times \chi_0$ with a two-dimensional nullspace corresponding to each χ_0 spanned by $\{\psi_1, \psi_1^*\}$. Therefore taken together on χ dim $N(\mathbf{I} - \Lambda\mathbf{K}) = 4$. □

With Proposition 2.2 we may decompose the space χ as $\chi = N(\mathbf{I} - \Lambda\mathbf{K}) \oplus R(\mathbf{I} - \Lambda\mathbf{K})$ and we define a projection operator $\mathbf{P} = \operatorname{diag}[\mathbf{P}_1, \mathbf{P}_1]$ where

$$\mathbf{P}_1 f = \psi_1(t)\langle \psi_1, f\rangle + \psi_1^*(t)\langle \psi_1^*, f\rangle.$$

$\langle \cdot, \cdot \rangle$ is a functional pairing on χ_0

$$\langle f, g\rangle = \int_0^{2\pi} f^*(\tau) g(\tau)\, d\tau.$$

It follows that $R(\mathbf{I} - \Lambda\mathbf{K}) = \{f \in \chi : \mathbf{P}f = 0\}$. Now the function ϕ is split into two parts

$$\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

where $u = \mathbf{P}\phi$ and $v = [\mathbf{I} - \mathbf{P}]\phi$ and we define the mapping from $N(\mathbf{I} - \Lambda\mathbf{K}) \to \mathrm{C}$ by

$$\alpha_j = \langle \psi_1, \phi_j \rangle \quad \text{for } j = 1, 2. \tag{2.5}$$

The equations may now be decomposed in Lemma 2.3.

LEMMA 2.3. *The function $\phi = u + v$ is a 2π periodic solution of $\Phi = 0$ if and only if*

$$[\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} h_1(u_1 + v_1, u_2 + v_2, t) \\ h_2(u_1 + v_1, u_2 + v_2, t) \end{pmatrix} \tag{2.6}$$

and

$$\mathbf{P}\begin{pmatrix} h_1(u_1 + v_1, u_2 + v_2, t) \\ h_2(u_1 + v_1, u_2 + v_2, t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{2.7}$$

where

$$\begin{aligned} h_1(\phi_1, \phi_2, t) = {} & (\lambda - 1)\mathbf{K}\phi_1 + \lambda(\sin \mathbf{K}\phi_1 - \mathbf{K}\phi_1) + 2\sqrt{2}\mu_1 \cos(t + \zeta) \cos \mathbf{K}\phi_1 \\ & - \frac{m}{1 - \sigma} \sin \mathbf{K}\phi_1 [\phi_2 \sin \mathbf{K}\phi_2 - (\mathbf{K}_t\phi_2)^2 \cos \mathbf{K}\phi_2], \end{aligned} \tag{2.8a}$$

$$\begin{aligned} h_2(\phi_1, \phi_2, t) = {} & (\lambda - 1 - \sigma)\mathbf{K}\phi_2 + \lambda(1 - \sigma)(\sin \mathbf{K}\phi_2 - \mathbf{K}\phi_2) + 2\sqrt{2}\mu_2 \cos t \cos \mathbf{K}\phi_2 \\ & - (\lambda - 1)\sigma\mathbf{K}\phi_2 - (1 - \sigma) \sin \mathbf{K}\phi_2 [\phi_1 \sin \mathbf{K}\phi_1 - (\mathbf{K}_t\phi_1)^2 \cos \mathbf{K}\phi_1] \end{aligned} \tag{2.8b}$$

and $[\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]$ is an isomorphism from $N(\mathbf{I} - \Lambda\mathbf{K})^\perp \to R(\mathbf{I} - \Lambda\mathbf{K})$.

Proof. The pair of equations (2.2) has simply been rearranged into a linear and a nonlinear part $h : \chi \to \chi$ that satisfies $D_\phi h(0, 0, t) = 0$ where D_ϕ is the Fréchet derivative. By the Fredholm alternative the equation $[\mathbf{I} - \Lambda\mathbf{K}] = h$ has a solution only if

$h \in R(\mathbf{I}-\Lambda\mathbf{K})$ or equally $h \in N(\mathbf{I}-\Lambda\mathbf{K})^{\perp}$. This yields (2.7). Equation (2.6) is found by projecting (2.2) along $N(\mathbf{I}-\Lambda\mathbf{K})$ onto its complement. The restriction of $[\mathbf{I}-\Lambda\mathbf{K}]$ to $N(\mathbf{I}-\Lambda\mathbf{K})^{\perp}$ is then an isomorphism $[\mathbf{I}-\Lambda\mathbf{K}+\mathbf{P}]: N(\mathbf{I}-\Lambda\mathbf{K})^{\perp} \to R(\mathbf{I}-\Lambda\mathbf{K})$. □

Properties of (2.6) are established in Lemma 2.4.

LEMMA 2.4. *There exists a solution* $v(t; \alpha, \lambda-1, \sigma, \mu)$ *of* (2.6) *satisfying* (a) $v(t; 0, 0, 0, 0) = 0$, *and* (b) for $|\alpha|, |\lambda-1|, |\sigma|$ *bounded away from zero the solution* v *of* (2.6) *is also bounded away from zero but tends to zero with its last four arguments.*

The mapping $h: \chi \to \chi$ *defined in* (2.8) *satisfies the following inequalities*:

$$\|h\|_{\chi} = \max\{\|h_1\|, \|h_2\|\} \leqq c_1(\alpha, \lambda-1, \sigma, \mu) \tag{2.9}$$

and

$$\max_{t=1,2}\left\{\sum_{j=1}^{2} \|(Dh_i)_{v_j}\|\right\} \leqq c_2(\alpha, \lambda-1, \sigma, \mu) \tag{2.10}$$

where c_1, c_2 *are positive bounded functions that tend to zero with their arguments.*

Proof. When $\alpha = \lambda - 1 = \sigma = \mu = 0$ v satisfies

$$[\mathbf{I}-\Lambda\mathbf{K}+\mathbf{P}]\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \sin \mathbf{K}v_1 - \mathbf{K}v_1 - m \sin \mathbf{K}v_1[v_2 \sin \mathbf{K}v_2 - (\mathbf{K}_t v_2)^2 \cos \mathbf{K}v_2] \\ \sin \mathbf{K}v_2 - \mathbf{K}v_2 - \sin \mathbf{K}v_2[v_1 \sin \mathbf{K}v_1 - (\mathbf{K}_t v_1)^2 \cos \mathbf{K}v_1] \end{pmatrix}. \tag{2.11}$$

The Fréchet derivative of (2.11) is $[\mathbf{I}-\Lambda\mathbf{K}+\mathbf{P}]$ which was established in Lemma 2.3 to be an isomorphism from $N(\mathbf{I}-\Lambda\mathbf{K})^{\perp} \to R(\mathbf{I}-\Lambda\mathbf{K})$. By the implicit function theorem there exists a unique $v \in \chi$ in a neighborhood of the origin. But note that $v = 0$ satisfies (2.11). It follows that $v = 0$ is the only solution *connected* to the origin and (a) follows.

When $\alpha, \lambda-1, \sigma, \mu$ are nonzero $v = 0$ is clearly *not* a solution of (2.6). It follows that $v \neq 0$ when $\alpha, \lambda-1, \sigma, \mu$ are nonzero but by (a) v must tend to zero with its last four arguments.

The inequalities will now be proved. From (2.8) we have that

$$\begin{aligned} \|h_1\| \leqq |\lambda-1|\,\|\mathbf{K}\|\,\|u_1+v_1\| + 2\sqrt{2}|\mu_1| + |\lambda|\,\|\sin \mathbf{K}(u_1+v_1) - \mathbf{K}(u_1+v_1)\| \\ + \left|\frac{m}{1-\sigma}\right|[\|u_2+v_2\| + \|\mathbf{K}_t\|^2\|u_2+v_2\|^2] \end{aligned} \tag{2.12}$$

and $\|h_2\|$ has a similar expression. Now $\|u_j\| \leqq 2|\alpha_j|$ for $j = 1, 2$, $\|\mathbf{K}\|$ is bounded and by the first part of this lemma $\|v\|_{\chi}$ tends to zero with its arguments. Therefore $\|h_1\|, \|h_2\|$ are bounded by functions of $\alpha, \lambda-1, \sigma, \mu$ which tend to zero with their arguments. Choosing the maximum of these we have c_1 in (2.9).

The Fréchet derivative of h_1 with respect to v_1 is

$$(Dh_1)_{v_1} = \left[(\lambda-1) + \lambda[\cos \mathbf{K}\phi_1 - \mathbf{I}] - 2\sqrt{2}\mu_1 \cos(t+\zeta) \sin \mathbf{K}\phi_1 - \frac{m}{1-\sigma} \cos \mathbf{K}\phi_1[\phi_2 \sin \mathbf{K}\phi_2 - (\mathbf{K}_t\phi_2)^2 \cos \mathbf{K}\phi_2]\right]\mathbf{K}.$$

Therefore

$$\|(Dh_1)_{v_1}\| \leqq \|\mathbf{K}\|\left[|\lambda-1| + |\lambda|\,\|\cos \mathbf{K}\phi_1 - \mathbf{I}\| + 2\sqrt{2}|\mu_1| + \frac{m}{1-\sigma}[\|\phi_2\| + \|\mathbf{K}_t\|^2\|\phi_2\|^2]\right]. \tag{2.13}$$

However, by part (a) we know that $\|\phi\|_\chi \to 0$ with $\alpha, \lambda - 1, \sigma, \mu$. Therefore the right-hand side of (2.13) is less than some number dependent on $\alpha, \lambda - 1, \sigma, \mu$ and going to zero with $\alpha, \lambda - 1, \sigma, \mu$. Repeating this argument for each of the other Fréchet derivatives of h_i and choosing the maximum of the bounds as c_2 we have (2.10). □

With Lemmas 2.3 and 2.4 the bifurcation equations are constructed in Theorem 1.

THEOREM 1. *There are positive constants $\varepsilon_0 > 0$, $\lambda_0 > 0$, $\sigma_0 > 0$, $\mu_0 > 0$ with $\|u\|_\chi \leqq \varepsilon_0$, $|\lambda - 1| \leqq \lambda_0$, $|\sigma| \leqq \sigma_0$, and $|\mu| \leqq \mu_0$ such that there exists a unique $\hat{v} \in \chi$ with $v = \hat{v}(t; \alpha, \lambda - 1, \sigma, \mu)$ which satisfies* (2.6). *$\hat{v}$ has a continuous first derivative with respect to its arguments and $\hat{v}(t; 0, 0, 0, 0) = 0$. Substitution of this unique v into* (2.7) *generates the bifurcation equations*

$$(\lambda - 1)\alpha_1 - \frac{1}{4\pi}|\alpha_1|^2\alpha_1 + 2\sqrt{\pi}\mu_1 e^{i\zeta} + \frac{1}{\pi}m\alpha_2^2\alpha_1^* + r_1(\alpha, \alpha^*, \lambda - 1, \sigma, \mu) = 0,$$

$$(\lambda - 1 - \sigma)\alpha_2 - \frac{1}{4\pi}|\alpha_2|^2\alpha_2 + 2\sqrt{\pi}\mu_2 - \frac{1}{\pi}\alpha_1^2\alpha_2^* + r_2(\alpha, \alpha^*, \lambda - 1, \sigma, \mu) = 0$$

and

$$\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix} = \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\psi_1(t) + \begin{pmatrix}\alpha_1^*\\ \alpha_2^*\end{pmatrix}\psi_1^*(t) + \begin{pmatrix}\hat{v}(t; \alpha, \lambda - 1, \sigma, \mu)\\ \hat{v}(t; \alpha, \lambda - 1, \sigma, \mu)\end{pmatrix}$$

is a 2π periodic solution of (2.2) *if and only if the bifurcation equations are nondegenerate. The remainder term $r = (r_1, r_2)$ satisfies*

$$\|r\|_\chi = 0\{(|\alpha_1| + |\alpha_2|)^2|\mu_1| + (|\lambda - 1| + |\sigma|)(|\alpha_1| + |\alpha_2|) + (|\alpha_1| + |\alpha_2|)^3\}$$

as $(|\alpha_1|, |\alpha_2|, |\lambda - 1|, |\sigma|, |\mu|) \to 0$.

Proof. $v = [\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]^{-1}h$ is a mapping from the ball $\|v\|_\chi \leqq v_0$ to itself when $c_1 \leqq 3v_0/7$. In addition $v = [\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]^{-1}h$ is a uniform contraction when $c_2 < 3/7$. As c_1, c_2 may be made arbitrarily small by proper choice of λ_0, ε_0, σ_0, and μ_0 the existence, and properties, of $\phi = u + \hat{v}$ are a consequence of the uniform contraction mapping theorem. The function $\hat{v}$ may be obtained as the limit of the sequence $\{v_n\}$ with $v_{n+1} = [\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]^{-1}h(v_n, t)$ for $n = 0, 1, \cdots$ with $v_0 = 0$. Using the first approximation we arrive at the bifurcation equations with the given error bounds. □

As the bifurcation equations are nonanalytic they will be analyzed by decomposing the amplitudes into $\alpha_j = \rho_j e^{i\xi_j}$ for $j = 1, 2$. This yields

$$(\lambda - 1)\rho_1 - \frac{1}{4\pi}\rho_1^3 - \frac{m}{\pi}\rho_2^2\rho_1 \cos 2(\xi_2 - \xi_1) + 2\sqrt{\pi}\mu_1 \cos(\xi_1 - \zeta) + \operatorname{Re}[r_1 e^{-i\xi_1}] = 0, \tag{2.14a}$$

$$(\lambda - 1 - \sigma)\rho_2 - \frac{1}{4\pi}\rho_2^3 - \frac{1}{\pi}\rho_1^2\rho_2 \cos 2(\xi_2 - \xi_1) + 2\sqrt{\pi}\mu_2 \cos \xi_2 + \operatorname{Re}[r_2 e^{-i\xi_2}] = 0, \tag{2.14b}$$

$$-\frac{m}{\pi}\rho_1\rho_2^2 \sin 2(\xi_2 - \xi_1) - 2\sqrt{\pi}\mu_1 \sin(\xi_1 - \zeta) + \operatorname{Im}[r_1 e^{-i\xi_1}] = 0, \tag{2.15a}$$

$$\frac{1}{\pi}\rho_1^2\rho_2 \sin 2(\xi_2 - \xi_1) - 2\sqrt{\pi}\mu_2 \sin \xi_2 + \operatorname{Im}[r_2 e^{-i\xi_2}] = 0. \tag{2.15b}$$

These equations are the basis of the qualitative behavior of the nondegenerate periodic solutions for the orthogonal planar pendulum. It provides a family of initial conditions and relative phases that result in periodic orbits. It is these equations that will be analyzed in the next two sections.

3. Autonomous oscillations and $Z_2 \oplus Z_2$ symmetry. Treating the phase difference $\xi_2 - \xi_1$ as a parameter the four bifurcation equations may be thought of as functions of $\rho = (\rho_1, \rho_2)$ and $\kappa = (\lambda, \xi_2 - \xi_1, \sigma, \mu)$, with $\mu = 0$,

$$f_1(\rho; \kappa) = \rho_1\left[\lambda - 1 - \frac{1}{4\pi}\rho_1^2 - \frac{m}{\pi}\cos 2(\xi_2 - \xi_1)\rho_2^2\right] + 0_5, \tag{3.1a}$$

$$f_2(\rho; \kappa) = \rho_2\left[\lambda - 1 - \sigma - \frac{1}{4\pi}\rho_2^2 - \frac{1}{\pi}\cos 2(\xi_2 - \xi_1)\rho_1^2\right] + 0_5, \tag{3.1b}$$

$$g_1(\rho; \kappa) = -\frac{m}{\pi}\rho_1\rho_2^2 \sin 2(\xi_2 - \xi_1) + 0_5, \tag{3.2a}$$

$$g_2(\rho; \kappa) = -\frac{1}{\pi}\rho_1^2\rho_2 \sin 2(\xi_2 - \xi_1) + 0_5 \tag{3.2b}$$

where 0_5 contains terms of fifth order and higher. Equations (3.1) are analogous to the normal form for the double cusp with symmetry [11, p. 219]. In the absence of (3.2) the function f when nondegenerate is $Z_2 \oplus Z_2$ equivalent to

$$f_1(x, y; \kappa) = x[x^2 + 4m\varepsilon_3 y^2 - (\lambda - 1)] = 0, \tag{3.3a}$$

$$f_2(x, y; \kappa) = y[4\varepsilon_3 x^2 + y^2 - (\lambda - 1 - \sigma)] = 0 \tag{3.3b}$$

where $\varepsilon_3 = \cos 2(\xi_2 - \xi_1)$. However, the presence of g and its coupling to f through the phase make the equivalence formal only. We have no rigorous basis for neglecting 0_5 in all four equations and therefore proceed on a formal basis by neglecting 0_5. This implies that $\sin 2(\xi_2 - \xi_1) = 0$ when $\rho_1\rho_2 \neq 0$; therefore we study solutions of (3.3) with $\varepsilon_3 = \pm 1$. The set (3.3) is the basic normal form for problems with $Z_2 \oplus Z_2$ symmetry and has been studied in much detail by Golubitsky and Schaeffer [12, Chap. X], from which we have the following proposition.

PROPOSITION 3.1. *The bifurcation problem f in* (3.3) *is degenerate when $m = \frac{1}{4}$ and $\varepsilon_3 = 1$ or whem $m = \frac{1}{16}$ and $\varepsilon_3 = \pm 1$.*

Proof. The equations become dependent at these values of ε_3 and m. The complete proof is given in [12, p. 423].

The implication is that the system has families of periodic solutions at admissible values of m and ε_3 other than these critical values. The points of degeneracy point towards more complex solutions. In § 6 some examples will be given to suggest what might occur in the system at these points of degeneracy.

For values of m and ε_3 other than the degenerate points, the periodic solutions corresponding to (3.3) are

$$\text{I:} \quad x = \sqrt{\lambda - 1}; \qquad y = 0,$$

$$\text{II:} \quad x = 0; \qquad y = \sqrt{\lambda - 1 - \sigma},$$

$$\text{III:} \quad x^2 = \frac{1 - 4m\varepsilon_3}{1 - 16m}\left[\lambda - 1 + \frac{4m\varepsilon_3}{1 - 4m\varepsilon_3}\sigma\right],$$

$$y^2 = \frac{1 - 4\varepsilon_3}{1 - 16m}\left[\lambda - 1 - \frac{\sigma}{1 - 4\varepsilon_3}\right]$$

and only nonnegative values of x, y are admissible. Solution sets I and II are pure mode pitchforks and correspond to the Lyapunov families emitted by the bifurcation points $\lambda = 1$ and $\lambda = 1 + \sigma$. The mixed mode solutions are divided into five sets separated by the points of degeneracy. For $\varepsilon_3 = +1$; $0 < m < \frac{1}{16}$, $\frac{1}{16} < m < \frac{1}{4}$, and $\frac{1}{4} < m < 1$ and for

$\varepsilon_3 = -1$; $0 < m < \frac{1}{16}$ and $\frac{1}{16} < m < 1$. Solution types for each of these regions have been classified by Golubitsky and Schaeffer [12, Chap. X]. The parameter σ which perturbs the 1:1 resonance plays the role of the universal unfolding for this normal form. The unfolding results in a secondary branching of the periodic solutions.

Although there is a translation invariance with respect to time, the requirement that $\sin 2(\xi_2 - \xi_1) = 0$ restricts the phase between the two pendulums. $\varepsilon_3 = 1$ corresponds to in phase motions and $\varepsilon_3 = -1$ corresponds to a phase difference of $\pm\pi/2$ between the two pendulums.

Figures 1-4 show examples of the multiplicity and nature of the autonomous periodic solutions in a neighborhood of the 1:1 resonance. Figure 1 contains in-phase and out-of-phase solutions at 1:1 resonance with $m = \frac{1}{2}$, and Fig. 2 is an unfolding of Fig. 1 with $\sigma = \frac{1}{2}$. Figure 3 shows a neighborhood of the 1:1 resonance for $m = \frac{1}{7}$ and $\sigma = -\frac{3}{5}$ with both in-phase and out-of-phase solutions. Note that there is a global connection between the two Lyapunov families. Figure 4 is similar with parameters $m = \frac{1}{20}$ and $\sigma = -\frac{3}{5}$ with in-phase and out-of-phase solutions. In Figs. 1-4 the in-phase coupled solutions are indicated by +1 and the out-of-phase coupled solutions are indicated by −1.

These figures are not bifurcation diagrams in the usual sense, although the terminology applies well, due to the fact that λ corresponds to the period of the oscillations. The figures are essentially maps of initial displacements and relative phase between the two modes such that periodic solutions result. A stability analysis of the solutions is carried out in § 5 using Floquet theory. That analysis shows that the in-phase motions, $\varepsilon_3 = +1$, are stable for all m such that $m > \frac{1}{7}$ while the out-of-phase coupled motions, $\varepsilon_3 = -1$, are stable for all m (the degenerate points excepted). The stability of the solutions implies that there is no exchange of energy between the two degrees of freedom during the motion.

4. Symmetry breaking due to forced oscillations. When the parameter μ is included in the analysis the bifurcation equations for ρ_1, ρ_2 no longer commute with the action of $\mathbf{Z}_2 \oplus \mathbf{Z}_2$. The symmetry is broken and μ provides an unfolding of the normal form. Although the Γ-codimension (the dimension of the unfolding that preserves the $Z_2 \oplus Z_2$ symmetry) is three (two modal parameters (m, ε_3) and σ) the *contact* codimension which allows for the dissolution of the $Z_2 \oplus Z_2$ symmetry is 16 [11, p. 219]. Consequently the addition of the parameters μ_1 and μ_2 is by no means a universal unfolding but the parameters have been introduced naturally and correspond to a measure of the forcing of the dynamical system. The addition of more parameters and further study is necessary before a complete picture of the forced oscillations may be formed. After neglecting terms of fifth order and higher, and performing a smooth change of variables, the bifurcation equations are

$$x(x^2 + 4m\varepsilon_3 y^2 - (\lambda - 1)) - \mu_1 \cos(\xi_1 - \zeta) = 0, \tag{4.1a}$$

$$y(4\varepsilon_3 x^2 + y^2 - (\lambda - 1 - \sigma)) - \mu_2 \cos \xi_2 = 0 \tag{4.1b}$$

with the constraints,

$$\mu_1 \sin(\xi_1 - \zeta) + 4mxy^2 \sin 2(\xi_2 - \xi_1) = 0, \tag{4.2a}$$

$$\mu_2 \sin \xi_2 - 4x^2 y \sin 2(\xi_2 - \xi_1) = 0. \tag{4.2b}$$

Equations (4.2) clearly include the special case

$$\sin(\xi_1 - \zeta) = \sin \xi_2 = 0. \tag{4.3}$$

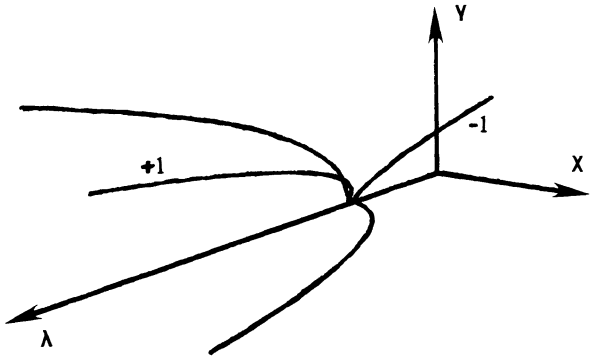


FIG. 1. *Bifurcation diagram for autonomous periodic solutions at* 1:1 *resonance with* $m=\frac{1}{2}$ *and* $\sigma=0$. $+1$ *corresponds to an in-phase mixed mode and* -1 *corresponds to an out-of-phase mixed mode.*

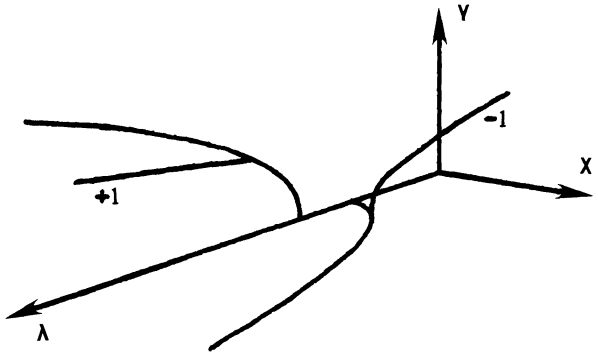


FIG. 2. *Same parameters as Fig.* 1 *but with* $\sigma=\frac{1}{2}$.

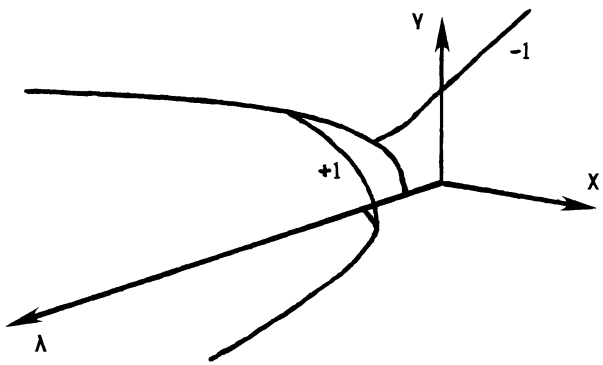


FIG. 3. *Bifurcation diagram for autonomous periodic solutions with* $m=\frac{1}{7}$ *and* $\sigma=-\frac{3}{5}$.

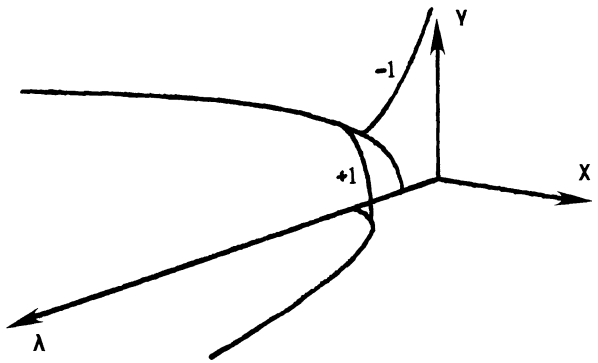


FIG. 4. *Bifurcation diagram for autonomous periodic solutions with* $m=\frac{1}{20}$ *and* $\sigma=-\frac{3}{5}$.

From this subclass of solutions some examples will be shown. We note however that there may be other families of periodic solutions with $\sin(\xi_1-\zeta)\neq 0$ and $\sin\xi_2\neq 0$. With the restriction (4.3) we have

$$f_1(x, y; \kappa) = x[x^2+4m\varepsilon_3 y^2-(\lambda-1)]-\mu_1\varepsilon_1 = 0, \tag{4.4a}$$

$$f_2(x, y; \kappa) = y[4\varepsilon_3 x^2+y^2-(\lambda-1-\sigma)]-\mu_2\varepsilon_3 = 0, \tag{4.4b}$$

with $\kappa=(\lambda, m, \sigma, \mu, \varepsilon)$ and $\varepsilon=(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ where

$$\varepsilon_1 = \cos(\xi_1-\zeta) = \pm 1,$$

$$\varepsilon_2 = \cos\xi_2 = \pm 1,$$

$$\varepsilon_3 = \cos 2(\xi_2-\xi_1) = \pm 1.$$

Particular solution sets for (4.4) are illustrated in Figs. 5-9.

In Fig. 5 $\mu_1=0$, $\zeta=0$, $m=\frac{1}{2}$, $\sigma=0$, and $\zeta=0$ forces $\varepsilon_3=+1$. This is an unfolding of Fig. 1. There are two branches with $x=0$. At $x=0$, $y=(\mu_2/2)^{1/3}$, $\lambda=1+3(\mu_2/2)^{2/3}$ there is a bifurcation point for a branch with $x\neq 0$. In addition there is an isolated branch with $x\neq 0$. Solid lines correspond to a stable branch and dashed lines correspond to an unstable branch. There is no qualitative difference when σ is nonzero here as indicated in Fig. 6. Figure 6 has the same parameters as Fig. 5 but with $\sigma>0$.

In Fig. 7 $\mu_1\neq 0$ but $\mu_2=\sqrt{3}\mu_1$, $\xi=0$, $m=\frac{1}{2}$, $\sigma=0$, and $\zeta=0$ requires that $\varepsilon_3=+1$. Here there are two branches satisfying $y=\sqrt{3}x$. On the lower of these there is a limit point at $x=(\mu_1/14)^{1/3}$, $\lambda=1+21(\mu_1/14)^{2/3}$. The bifurcation point, at $x=(\mu_1/6)^{1/3}$, $\lambda=1+(\mu_1/6)^{2/3}$, on the upper symmetric branch is a pitchfork bifurcation and emits two *unstable* branches with $y\neq\sqrt{3}x$. The stability properties are indicated by dashed and solid lines. Figure 8 is similar to Fig. 7 but a phase is introduced in the forcing function; $\zeta=\pi/2$, and $\mu_2=\sqrt{5/3}\mu_1$. This generates two branches with $y=\sqrt{5/3}x$. On the lower of these there is a limit point at $x=(3\mu_1/14)^{1/3}$, $\lambda=1+\frac{23}{20}(4\mu_2/3)^{2/3}$. The bifurcation point at $x=(\mu_1/10)^{1/3}$, $\lambda=1+\frac{23}{3}(\mu_1/10)^{2/3}$ is a pitchfork bifurcation and it emits one locally stable and one locally unstable branch with $y\neq\sqrt{5/3}x$.

Figure 9 is a σ-unfolding of Fig. 7. The parameters are the same as those of Fig. 7 but with $\sigma=.005$. In this situation σ splits the bifurcation point on the upper symmetric branch producing a limit point. The results in Figs. 5-9 correspond to $m=\frac{1}{2}$ (the two masses are equal). μ-unfoldings for $\frac{1}{16}<m<\frac{1}{4}$ and $0<m<\frac{1}{16}$ should also provide interesting results.

5. Stability analysis of the periodic solutions. The stability of the periodic solutions found in the previous sections will be determined using Floquet theory. As the periodic orbits themselves are only known locally the stability properties will be determined locally using the Lyapunov-Schmidt method to determine the sign of the Floquet exponents in the neighborhood of a bifurcation point.

The Fréchet derivative of the coupled ordinary differential equations governing the motion of the orthogonal planar pendulum results in the linear system

$$[\mathbf{I}+\mathbf{A}_1(t)]\frac{d^2\theta'}{dt^2}+\mathbf{A}_2(t)\frac{d\theta'}{dt}+[\mathbf{I}+\mathbf{A}_3(t)]\theta' = 0 \tag{5.1}$$

where the 2×2 matrices $\mathbf{A}_j(t)$ for $j=1, 2, 3$ satisfy $\mathbf{A}_j(t+2\pi)=\mathbf{A}_j(t)$ and expressions for each of them are given in the Appendix.

From Floquet theory we know that (5.1) has a solution $\theta'(t)=\theta''(t)\,e^{\eta t}$ with the stipulation that $\theta''(t+2\pi)=\theta''(t)$ and η is the Floquet exponent that determines the stability of the periodic orbits. Sign $[\mathrm{Re}(\eta)]<0$ implies stability and sign $[\mathrm{Re}(\eta)]>0$ implies instability.

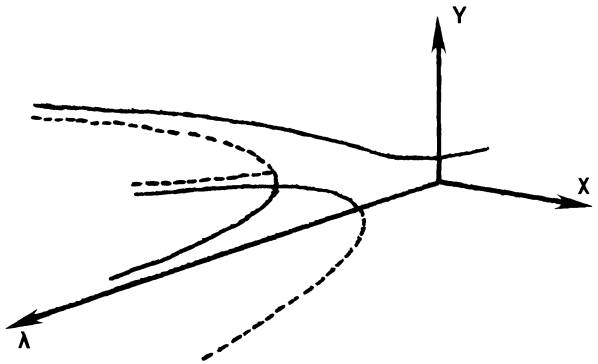


FIG. 5. *Bifurcation diagram for forced periodic solutions with one component forcing.* $m=\frac{1}{2}$, $\mu_1=0$, $\mu_2=\frac{1}{4}$, $\zeta=0$, *and* $\sigma=0$. *Stable branches are solid and unstable branches are dashed.*

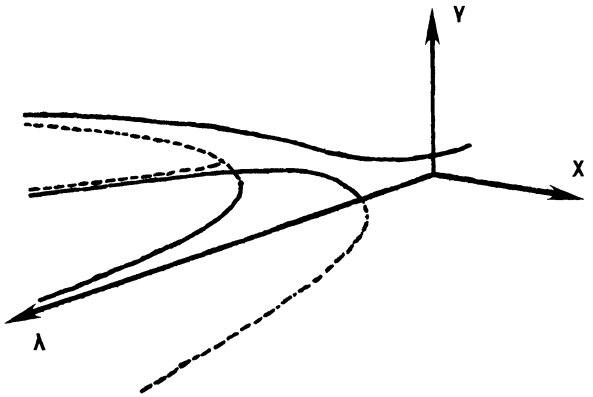


FIG. 6. *Same parameters as Fig. 5 but with* $\sigma=\frac{3}{10}$.

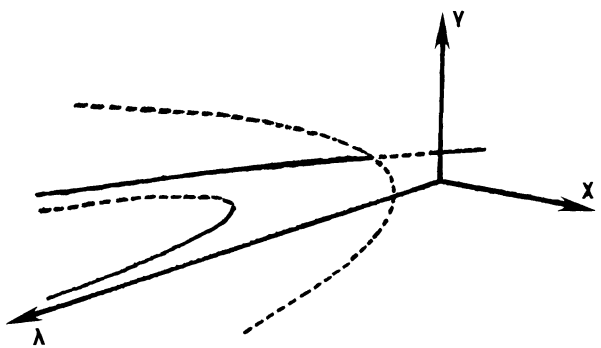


FIG. 7. *Bifurcation diagram for forced periodic solutions with two component forcing.* $m=\frac{1}{2}$, $\mu_1=\frac{1}{4}$, $\mu_2=\sqrt{3}\mu_1$, $\zeta=0$, *and* $\sigma=0$.

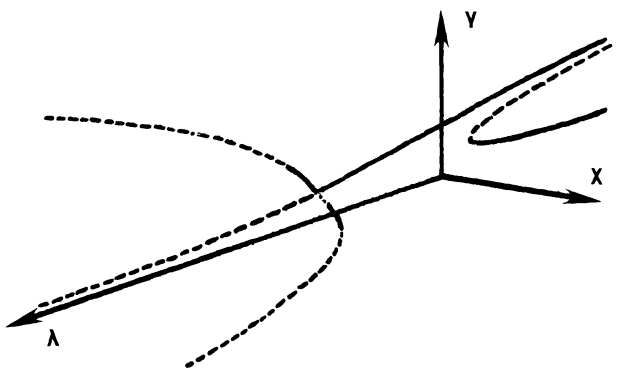


FIG. 8. *Bifurcation diagram for forced periodic solutions with nonzero* ζ. $m=\frac{1}{2}$, $\mu_1=\frac{1}{4}$, $\mu_2=\sqrt{\frac{5}{3}}\mu_1$, $\zeta=\pi/2$, *and* $\sigma=0$.

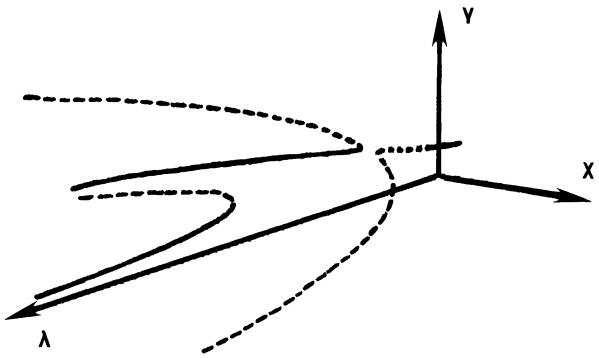


FIG. 9. *Effect of σ on two component forced oscillations. Same parameters as Fig. 7 but with $\sigma = .005$.*

The Lyapunov-Schmidt method will be used to estimate η near the origin. To simplify the analysis (5.1) is converted to an integral equation for $\zeta(t) = -\mathbf{K}\theta''$ where $\zeta = (\zeta_1, \zeta_2)$ and $\zeta(t) = \theta''$. After some manipulation the equation for ζ is

$$[\mathbf{I} - \Lambda\mathbf{K}]_\zeta = [-\mathbf{A}_1 + (2\eta(\mathbf{I} + \mathbf{A}_1) + \mathbf{A}_2)\mathbf{K}_t + (\eta^2(\mathbf{I} + \mathbf{A}_1) + \eta\mathbf{A}_2 + \mathbf{A}_3)\mathbf{K}]\zeta, \tag{5.2}$$

which is a mapping from $\chi \to \chi$ and terms are as previously defined. For (5.2) we have the following lemma.

LEMMA 5.1. *Equation* (5.2) *has a* 2π *periodic solution only if* $\zeta = a\cos t + b\sin t + \zeta'$, *where* $a, b \in R^2$, *satisfies the system of equations*

$$[\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]\zeta' = h(\zeta', t; a, b) \tag{5.3}$$

and

$$\mathbf{P}h = 0 \tag{5.4}$$

where h is given by the right-hand side of (5.2).

Proof. It was previously established that $[\mathbf{I} - \Lambda\mathbf{K}]$ is a Fredholm operator and $\chi = N(\mathbf{I} - \Lambda\mathbf{K}) \oplus R(\mathbf{I} - \Lambda\mathbf{K})$ and the Fredholm alternative applies. This requires that $h \in R(\mathbf{I} - \Lambda\mathbf{K})$ resulting in (5.4). The operator $[\mathbf{I} - \Lambda\mathbf{K} + \mathbf{P}]$ is a restriction of $[\mathbf{I} - \Lambda\mathbf{K}]$ to $N(\mathbf{I} - \Lambda\mathbf{K})^\perp$. □

Is is now straightforward to apply the implicit function theorem to (5.3) resulting in a unique ζ' in a neighborhood of the origin. Substitution of this solution into (5.4) will then result in the "bifurcation-stability" equations,

$$(\lambda - 1)A_1 + 2i\eta A_1 - \frac{2m}{\pi}\overline{\alpha_1}\alpha_2 A_2 - \frac{1}{4\pi}\alpha_1^2\overline{A_1} - \frac{1}{2\pi}|\alpha_1|^2 A_1 - \frac{m}{\pi}\alpha_2^2\overline{A_1} + r_1 = 0, \tag{5.5a}$$

$$\begin{aligned}(\lambda - 1 - \sigma)A_2 + 2i\eta A_2 - \frac{2}{\pi}\alpha_1\overline{\alpha_2}A_1 - \frac{1}{\pi}\alpha_1^2\overline{A_2} \\ - \frac{1}{4\pi}[\alpha_2^2\overline{A_2} + 2|\alpha_2|^2 A_2] + r_2 = 0\end{aligned} \tag{5.5b}$$

where r_1 and r_2 are higher order terms which satisfy

$$\|r\| = O\{(|\alpha_1| + |\alpha_2|)^2(|\lambda - 1| + |\sigma| + (|\alpha_1| + |\alpha_2|)^2) + |\sigma|\,|\lambda - 1| + |\eta|^2\}$$

and $A_1=(a_1-ib_1)/\sqrt{2}$ and $A_2=(a_2-ib_2)/\sqrt{2}$. If we formally neglect the higher order terms and expand the set (5.5) into four real homogeneous equations, then a solution exists only if the determinant of the coefficient matrix is zero. This yields a fourth-order algebraic equation for the Floquet exponents,

$$\Delta(\eta)=\eta^4-2b\eta^2+|J_1|\cdot|J_2|=0$$

where

$$b=-\frac{\mu_1}{2}\varepsilon_1[2x+\mu_1\varepsilon_1/x^2-8my^2\varepsilon_3/x]-\frac{\mu_2}{2}\varepsilon_2[2y+\mu_2\varepsilon_2/y^2-8x^2\varepsilon_3/y]$$
$$-8x^2y^2[(8-\varepsilon_3)m-\varepsilon_3]$$

and

$$J_1=\begin{pmatrix}2x+\mu_1\varepsilon_1/x^2 & 8my\varepsilon_3\\ 8x\varepsilon_3 & 2y+\mu_2\varepsilon_2/y^2\end{pmatrix},$$
$$J_2=\begin{pmatrix}-\mu_1\varepsilon_1+8mxy^2\varepsilon_3 & -8mxy^2\varepsilon_3\\ -8x^2y\varepsilon_3 & -\mu_2\varepsilon_2+8x^2y\varepsilon_3\end{pmatrix}.$$

When $\mu=0$, $|J_1|\cdot|J_2|=0$ and stability is assured if $b<0$, that is when $(8-\varepsilon_3)m-\varepsilon_3>0$. This yields the result stated in § 3 for stability of autonomous oscillations. Out of phase oscillations ($\varepsilon_3=-1$) are stable for all m and in phase oscillations ($\varepsilon_3=+1$) are stable for $m>\frac{1}{7}$.

When $\mu\neq 0$, the important points are those where $|J_1|\cdot|J_2|=0$. These correspond to points where there is an exchange of stability. Rather than a complete analysis, stability properties corresponding to Figs. 5-9 will be given. For Figs. 5 and 6 $\mu_1=0$ and $|J_1|=4x[(1-16m)y+\mu_2\varepsilon_3/2y^2]$. If $x\neq 0$, then $|J_1|=0$ if

$$y=\left(\frac{\mu_2\varepsilon_2}{2(16m-1)}\right)^{1/3}.$$

In Fig. 5 this corresponds to the limit point for the $x\neq 0$ branch. With $\mu_1=0$ $|J_2|=-8mxy^2\mu_2$, which provides no additional information. When $x=0$ a stability exchange occurs when $b=0$. This occurs when $y^3=(-\mu_2\varepsilon_2/2)$. This is the limit point on the lower $x=0$ branch. A similar result holds for $\sigma\neq 0$ with the location of the critical points perturbed. The results are indicated in Figs. 5 and 6.

For Figs. 7 and 8 a general analysis is more difficult but the following information may be obtained. Along the symmetric branches, where $y=kx$, $|J_1|=0$ when

$$x=\left[\frac{\mu_1}{28}\left(\frac{k^2-1}{k^2}\right)\left[\sqrt{1+\frac{28k^2}{(k^2+1)^2}}+\varepsilon_1\right]\right]^{1/3} \tag{5.6}$$

where $k^2=3$ for Fig. 7 and $k^2=5/3$ for Fig. 8, and $|J_2|=0$ when

$$x=\left(\frac{\mu_1\varepsilon_1}{4\varepsilon_3(2+k^2)}\right)^{1/3}. \tag{5.7}$$

First, for Fig. 7, (5.6) results in a bifurcation point at $x=(\mu_1/6)^{1/3}$, $\lambda=1+(\mu_1/6)^{2/3}$ on the upper symmetric branch and a limit point at $x=(\mu_1/14)^{1/3}$, $\lambda=1+21(\mu_1/14)^{2/3}$ on the lower symmetric branch. $|J_2|=0$ when $x=(\mu_1/20)^{1/3}$, $\lambda=1-14(\mu_1/20)^{2/3}$. Therefore we have that the solutions along the upper symmetric branch are stable for $0<\lambda<1-14(\mu_1/20)^{2/3}$, unstable for $1-14(\mu_1/20)^{2/3}<\lambda<1+(\mu_1/6)^{2/3}$, and stable for $\lambda>1+(\mu_1/6)^{2/3}$. The unsymmetric branches emitted by the bifurcation point are

unstable. The anomaly here is that there are no *stable* periodic solutions for $1-14(\mu_1/20)^{2/3}<\lambda<1+(\mu_1/6)^{2/3}$. This is a small neighborhood of the linear resonance $\lambda=1$.

A similar analysis for Fig. 8 shows that the upper symmetric branch is stable for $\lambda>\lambda_b$ where λ_b is the bifurcation point, and there is an exchange of stability at the limit point on the lower symmetric branch. For the unsymmetric branch the stability properties are anomalous. For the $+x$ branch there is a small region of stable solutions, separated from the bifurcation point by a region of unstable solutions, and for the $-x$ branch the solutions are stable for a small distance and then they become unstable. A blow-up of the region around the bifurcation point is shown in Fig. 10. The reason for these stability regions is not clear.

6. Remarks. The basis of §§ 3 and 4 is that the bifurcation equations *when nondegenerate* correspond to periodic solutions. When the bifurcation equations are degenerate we expect other solutions to occur or more complex families of periodic solutions. To illustrate what may occur at degenerate points we consider the Hamiltonian of Hénon and Heiles [14] given in (2.3)-(2.4). Following the analysis in this paper the bifurcation equations for the Hénon-Heiles problem are:

$$x[x^2+py^2-(\lambda-1)]=0, \tag{6.1a}$$

$$y[qx^2+y^2-(\lambda-1)]=0 \tag{6.1b}$$

with

$$p=\frac{2}{5r}(1-\varepsilon_3/6), \tag{6.2a}$$

$$q=\tfrac{1}{5}(4+18r+3(2-r)\varepsilon_3) \tag{6.2b}$$

with $\varepsilon_3=\pm1$ and $r=d/c$. When $p\neq1$, $q\neq1$, and $pq\neq1$ the equations are nondegenerate and periodic solutions occur. This has been proved from a different point of view by Braun [2] and Kummer [18]. When $r=-\frac{1}{3}$ and $\varepsilon_3=1$ the normal form is degenerate; $q=1$. (This is analogous to the degeneracy $m=\frac{1}{4}$, $\varepsilon_3=1$ for the O-P pendulum.) The critical value $r=-\frac{1}{3}$ is precisely the Hamiltonian studied by Hénon and Heiles [14] and Lichtenberg and Lieberman [19]. Lichtenberg and Lieberman [19, pp. 46-50] show that regions of stochasticity occur in the phase space for this particular set of parameters.

This normal form is also similar to the normal form for the spherical pendulum studied by Miles [22]. Using the approach in this paper we find that the normal form

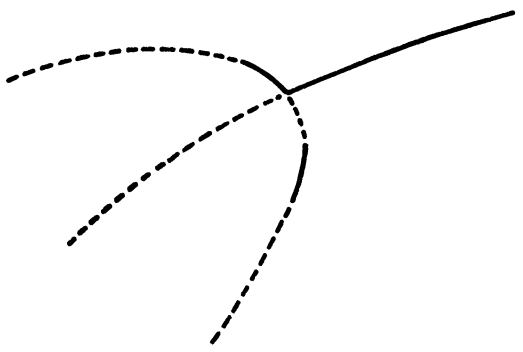


FIG. 10. *Blow-up of the neighborhood of the bifurcation point in Fig. 8 showing the regions of stability.*

for the spherical pendulum without forcing or damping is

$$x[x^2+(3\varepsilon_3-2)y^2-(\lambda-1)]=0,$$
$$y[(3\varepsilon_3-1)x^2+y^2-(\lambda-1)]=0,$$

which is clearly degenerate when $\varepsilon_3=1$. At this value the form is similar to the form for the O-P pendulum when $m=\frac{1}{4}$, $\varepsilon_3=1$, and the Hénon-Heiles form with $r=-\frac{1}{3}$. Miles [22] has shown that the spherical pendulum has quasi-periodic as well as chaotic solutions when damping and forcing are present.

The degeneracy corresponding to $m=\frac{1}{16}$ in the O-P pendulum is more severe. The equations are completely degenerate. However this is precisely the type of degeneracy studied by Erneux and Reiss [8] in their study of equilibrium solutions, and Erneux and Matkowsky [9]. Erneux and Matkowsky showed that a bifurcation from periodic solutions into quasi-periodic solutions may occur in the neighborhood of a degeneracy of this type in the context of the Hopf bifurcation.

Knobloch [16] has begun a systematic classification of degenerate normal forms for $O(2)$ symmetric Hopf bifurcation. The $O(2)$ symmetric normal form is a special case of the $Z_2\oplus Z_2$ symmetric normal form, therefore those results may also be applicable to the degenerate 1:1 resonance in Hamiltonian systems. His approach is to include the appropriate higher order terms and this leads to interesting new classes of periodic solutions.

Appendix. The matrices $\mathbf{A}_1(t)$, $\mathbf{A}_2(t)$, $\mathbf{A}_3(t)$ introduced in (5.1) are defined here.

$$\mathbf{A}_1(t)=\begin{pmatrix} 0 & \dfrac{m}{1-\sigma}\sin \mathbf{K}\phi_1 \sin \mathbf{K}\phi_2 \\ (1-\sigma)\sin \mathbf{K}\phi_1 \sin \mathbf{K}\phi_2 & 0 \end{pmatrix},$$

$$\mathbf{A}_2(t)=\begin{pmatrix} 0 & 2\dfrac{m}{1-\sigma}(\mathbf{K}_t\phi_2)\sin \mathbf{K}\phi_1 \cos \mathbf{K}\phi_2 \\ 2(1-\sigma)(\mathbf{K}_t\phi_1)\cos \mathbf{K}\phi_1 \sin \mathbf{K}\phi_2 & 0 \end{pmatrix},$$

$$\mathbf{A}_3(t)=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and

$$a_{11}=(\lambda-1)\cos \mathbf{K}\phi_1+[\cos \mathbf{K}\phi_1-1]-2\sqrt{2}\mu_1 \cos t \sin \mathbf{K}\phi_1$$
$$-\frac{m}{1-\sigma}\cos \mathbf{K}\phi_1[\phi_2 \sin \mathbf{K}\phi_2+(\mathbf{K}_t\phi_2)^2 \cos \mathbf{K}\phi_2],$$
$$a_{12}=-\frac{m}{1-\sigma}\sin \mathbf{K}\phi_1[\phi_2 \cos \mathbf{K}\phi_2+(\mathbf{K}_t\phi_2)^2 \sin \mathbf{K}\phi_2],$$
$$a_{21}=-(1-\sigma)\sin \mathbf{K}\phi_2[\phi_1 \cos \mathbf{K}\phi_1+(\mathbf{K}_t\phi_1)^2 \sin \mathbf{K}\phi_1],$$
$$a_{22}=[\lambda(1-\sigma)-1]\cos \mathbf{K}\phi_2+[\cos \mathbf{K}\sigma_2-1]-2\sqrt{2}\mu_2 \cos t \sin \mathbf{K}\phi_2$$
$$=-(1-\sigma)\cos \mathbf{K}\phi_2[\phi_1 \sin \mathbf{K}\phi_1-(\mathbf{K}_t\phi_1)^2 \cos \mathbf{K}\phi_1].$$

REFERENCES

[1] L. BAUER, H. B. KELLER, AND E. L. REISS, *Multiple eigenvalues lead to secondary bifurcation*, SIAM Rev., 17 (1975), pp. 101-122.

[2] M. BRAUN, *On the applicability of the third integral of motion*, J. Differential Equations, 13 (1973), pp. 300–318.

[3] T. J. BRIDGES, *On the secondary bifurcation of three-dimensional standing waves*, SIAM J. Appl. Math., 47 (1987), pp. 40–59.

[4] ———, *Secondary bifurcation and change of type for three-dimensional standing waves in finite depth*, J. Fluid Mech., 179 (1987), pp. 137–153.

[5] E. BUZANO, *Secondary bifurcations of a thin rod under axial compression*, SIAM J. Math. Anal., 17 (1986), pp. 312–321.

[6] S. CAPRINO, C. MAFFEI, AND P. NEGRINI, *Hopf bifurcation at* 1:1 *resonance*, Nonlinear Anal. Theoret. Meth. Appl., 8 (1984), pp. 1011–1032.

[7] S. N. CHOW, J. MALLET-PARET, AND J. YORKE, *Global Hopf bifurcation from a multiple eigenvalue*, Nonlinear Anal. Theory, Meth. Appl., 2 (1978), pp. 753–763.

[8] T. ERNEUX AND E. L. REISS, *Singular secondary bifurcation*, SIAM J. Appl. Math., 44 (1984), pp. 463–478.

[9] T. ERNEUX AND B. J. MATKOWSKY, *Quasi-periodic waves along a pulsating propagating front in a reaction diffusion system*, SIAM J. Appl. Math., 44 (1984), pp. 536–544.

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

[11] M. GOLUBITSKY AND D. SCHAEFFER, *Imperfect bifurcation in the presence of symmetry*, Comm. Math. Phys., 67 (1979), pp. 205–232.

[12] ———, *Singularities and Groups in Bifurcation Theory, Vol.* I, Springer-Verlag, New York, 1985.

[13] J. HALE, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.

[14] M. HÉNON AND C. HEILES, *The applicability of the third integral of motion; some numerical experiments*, Astronom. J., 69 (1964), pp. 73–79.

[15] H. KIELHÖFER, *Hopf bifurcation at multiple eigenvalues*, Arch. Rational Mech. Anal., 69 (1979), pp. 53–83.

[16] E. KNOBLOCH, *On the degenerate Hopf bifurcation with* $O(2)$ *symmetry*, in Multiparameter Bifurcation Theory, AMS Cont. Math. Series 56, 1986, pp. 193–201.

[17] G. A. KRIEGSMANN AND E. L. REISS, *New magnetohydrodynamic equilibria by secondary bifurcation*, Phys. Fluids, 21 (1978), pp. 258–263.

[18] M. KUMMER, *On resonant nonlinearly coupled oscillators with two equal frequencies*, Comm. Math. Phys., 48 (1976), pp. 53–79.

[19] A. J. LICHTENBERG AND M. A. LIEBERMAN, *Regular and Stochastic Motion*, Springer-Verlag, New York, 1983.

[20] S. B. MARGOLIS AND B. J. MATKOWSKY, *Flame propagation in channels: secondary bifurcation to quasi-periodic pulsations*, SIAM J. Appl. Math., 45 (1985), pp. 93–129.

[21] K. R. MEYER, *Bibliographic notes on generic bifurcation in Hamiltonian systems*, in Multiparameter Bifurcation Theory, AMS Cont. Math Series 56, 1986, pp. 373–381.

[22] J. MILES, *Resonant motion of a spherical pendulum*, Phys. D, 11 (1984), pp. 309–323.

[23] J. MOSER, *Periodic solutions near an equilibrium and a theorem by Alan Weinstein*, Comm. Pure Appl. Math., 29 (1976), pp. 727–747.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

STUDY OF A DOUBLY NONLINEAR HEAT EQUATION WITH NO GROWTH ASSUMPTIONS ON THE PARABOLIC TERM*

DOMINIQUE BLANCHARD† AND GILLES A. FRANCFORT†

Abstract. A doubly nonlinear equation with no growth assumptions on the parabolic term or on the heat flux is studied. Two existence and comparison results are established under different assumptions on the data. The technique uses truncation-penalization of the energy and energy estimates through convex conjugate functions.

Key words. heat equation, fast growing energy, nonlinear flux, energy estimates, convexity

AMS(MOS) subject classification. 35K55

Introduction. Doubly nonlinear evolution equations of the form

$$\frac{\partial b(u)}{\partial t} - \operatorname{div} A(\nabla u) = f \quad \text{on } \Omega \times (0, T),$$

$$u = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$b(u)|_{t=0} = b(u_0),$$

Ω bounded domain of $\mathbb{R}^N$

were first studied, to our knowledge, by Lions [8], Raviart [10], and Bamberger [2] in the case where

$$b(u) = |u|^{\alpha-2}u, \qquad A(w) = |w|^{p-2}w.$$

Grange and Mignot [7] address this problem in an abstract setting, namely

$$\frac{d}{dt}(Bu) + Au = f, \qquad Bu|_{t=0} = Bu_0,$$

where A and B denote the subdifferentials of the convex functions Φ and Ψ. The analysis developed in [7] is based on the essential restriction that Φ must be continuous on a Banach space V_1, and Ψ on a Banach space V_2, where V_1 is *densely and compactly embedded* in V_2. Power type nonlinearities are then restricted to satisfy

$$\frac{1}{\alpha} > \frac{1}{p} - \frac{1}{N}.$$

Furthermore A and B are assumed to be bounded on the bounded sets of V_1 and V_2 and Φ is assumed to be coercive.

Similar equations are also investigated with the help of semigroup techniques in L_1 (cf., e.g., Benilan [3]).

In this paper existence of a solution of semi-abstract equations of the form

$$\frac{\partial}{\partial t} b(u) - \operatorname{div} D\Phi(\nabla u) = f \quad \text{in } \Omega \times (0, T),$$

* Received by the editors March 19, 1987; accepted for publication (in revised form) January 25, 1988.
† Laboratoire Central des Ponts et Chaussées, 58 boulevard Lefebvre, 75732 Paris Cedex 15, France.

$$u = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$b(u)|_{t=0} = b(u_0),$$

is established in the following framework:

—Φ is a C^1 convex functional on $[L_q(\Omega)]^N$, $q > 1$, with $\Phi(w) \geqq \beta(\int_\Omega |w|^q\, dx)^{r/q}$, $r > 1$, $2N/(N+2) < q$.

—b is a locally Lipschitz monotone real-valued function.

Loosely speaking, there need not exist a Banach space V_2 on which b is the subdifferential of a convex *continuous* function Ψ, and, if it exists, V_1 need not be embedded in V_2. We are, for instance, in a position to solve the doubly nonlinear evolution equation with power type nonlinearities for any values of α and p (greater than one). Furthermore the function b may grow faster than any power function at infinity ($b(u) = e^{e^u}$, for example). In contrast, it need not be strictly increasing on any part of $\mathbb{R}$. Thus the evolution equation may become stationary in subdomain of $\Omega \times (0, T)$.

Similar results are given by Alt and Luckhaus [1] in a setting that includes equations of the form

$$\frac{db(u)}{\partial t} - \operatorname{div} A(\nabla u) = f,$$

where A is a monotone strongly elliptic operator on $\mathbb{R}^N$, i.e.,

$$(A(z) - A(z'), z - z')_{\mathbb{R}^N} \geqq \alpha |z - z'|^p.$$

Note that in the case when $A(w) = |w|^{p-2}w$, p is then restricted to be greater than or equal to two.

In Alt and Luckhaus [1], as well as in Grange and Mignot [7], the proof of the existence of a solution is based on a backward time difference scheme. Our method uses penalization through addition of a term of the form $\varepsilon(\partial u/\partial t)$ together with a truncation of the function b.

The detailed hypotheses on b, Φ, the initial condition $b(u_0)$, and the forcing term f are given in § 1, together with the existence results. The first result (Theorem 1) is concerned with forcing terms f in $W^{1,1}(0, T; L_2(\Omega))$ and initial conditions $b(u_0)$ in $L_2(\Omega)$ with u_0 in $W_0^{1,q}(\Omega)$. It states the existence of a solution u that also satisfies a maximum principle if f has a distinguished sign. The second result (Theorem 2) addresses the case of a forcing term f in $W^{1,1}(0, T; W^{-1,q'}(\Omega))(1/q + 1/q' = 1)$ and an initial condition $b(u_0)$ in $L_1^{\text{loc}}(\Omega) \cap W^{-1,q'}(\Omega)$ with u_0 in $W_0^{1,q}(\Omega)$.

Section 2 is devoted to the proof of Theorem 1 while § 3 addresses the proof of Theorem 2. The details of the different steps are briefly described at the end of § 1. It should be noted however that our proof of Theorem 2 (§ 3) is inspired by the Lemmas 1.8 and 1.9 of Alt and Luckhaus [1].

Throughout the paper, the notation $\|u\|_{m,s}$ denotes the usual Sobolev norm of u on $W^{m,s}(\Omega)$, where $W^{m,s}(\Omega)$ is the space of all $L_s(\Omega)$-functions with derivatives up to order m in $L_s(\Omega)$. Unless otherwise specified, the product $\langle\, , \rangle$ stands for the duality product between $W_0^{1,q}(\Omega)$ and $W^{-1,q'}(\Omega)$.

1. Assumptions and statement of the existence results. Let Ω be a bounded domain of $\mathbb{R}^N$ ($N \geqq 1$) with Lipschitz boundary $\partial\Omega$. Let q, r, α, and T be four real numbers

satisfying

$$1<q<+\infty,$$

$$q>\frac{2N}{N+2}, \qquad r>1, \tag{1}$$

$$\alpha>0, \qquad T>0.$$

Inequalities (1) result in the following *compact* imbeddings:

$$W_0^{1,q}(\Omega)\hookrightarrow L_2(\Omega)\hookrightarrow W^{-1,q'}(\Omega). \tag{2}$$

In (2) the space $W_0^{1,q}(\Omega)$ is the subspace of all $W^{1,q}(\Omega)$-functions with null traces, whereas q' is the conjugate of q, i.e., $1/q+1/q'=1$.

Let b be defined as a real-valued function of the real variable with the following properties:

$$b \text{ is locally Lipschitz,}$$

$$b \text{ is monotone increasing,} \tag{3}$$

$$b(0)=0.$$

Remark 1. The function b is *not* restricted by any growth assumption at infinity, *nor* is it assumed to be strictly increasing.

If Ψ denotes the primitive of b, i.e.,

$$\Psi(t)=\int_0^t b(s)\,ds,$$

Ψ is a positive C^1 convex function, and its convex conjugate function Ψ^*, defined as

$$\Psi^*(t)=\sup_{s\in\mathbb{R}}\{ts-\Psi(s)\},$$

satisfies, for every t of $\mathbb{R}$,

$$\Psi^*(t)\geqq 0, \qquad \Psi^*(b(t))=b(t)t-\Psi(t). \tag{4}$$

Let Φ be defined as a real-valued functional on $[L_q(\Omega)]^N$ with the following properties:

$$\Phi \text{ is } C^1,$$

$$\Phi \text{ is convex,}$$

$$D\Phi \text{ is bounded on the bounded sets of } [L_q(\Omega)]^N, \tag{5}$$

$$\Phi(0)=0,$$

$$\Phi(w)\geqq \alpha\|w\|_{0,q}^r \quad \text{for any } w \text{ in } [L_q(\Omega)]^N.$$

The remainder of this paper is devoted to the proof of the following theorems.

THEOREM 1. *Under the assumptions* (1), (3), *and* (5), *and if*

$$u_0\in W_0^{1,q}(\Omega), \qquad b(u_0)\in L_2(\Omega), \tag{6}$$

$$f\in W^{1,1}(0,T;L_2(\Omega)), \tag{7}$$

the problem

$$\begin{aligned} &\frac{\partial b(u)}{\partial t} - \operatorname{div} D\Phi(\nabla u) = f \quad \text{in } \Omega \times (0, T), \\ &u = 0 \quad \text{on } \partial\Omega \times (0, T), \\ &b(u)|_{t=0} = b(u_0), \end{aligned} \tag{8}$$

admits a solution u such that

$$u \in L_\infty(0, T; W_0^{1,q}(\Omega)), \tag{9}$$

$$b(u) \in L_\infty(0, T; L_2(\Omega)) \cap W^{1,\infty}(0, T; W^{-1,q'}(\Omega)). \tag{10}$$

The norm u in $L_\infty(0, T; W_0^{1,q}(\Omega))$ *is bounded above by a continuous function of* $\Phi(\nabla u_0)$ *and of the norm* $|||f|||$ *of f in* $W^{1,1}(0, T; W^{-1,q'}(\Omega))$.

Furthermore, if u_{01} *and* u_{02} *satisfy* (6); *while* f_1 *and* f_2 *satisfy* (7), *and if* $b(u_{01}) - b(u_{02})$ *is almost everywhere positive on* Ω *while* $f_1 - f_2$ *is almost everywhere positive on* $\Omega \times (0, T)$, *there exist a solution* u_1 *associated to* u_{01}, f_1 *and a solution* u_2 *associated to* u_{02}, f_2 *such that* $b(u_1) - b(u_2)$ *is almost everywhere positive on* $\Omega \times (0, T)$. □

THEOREM 2. *Under the assumption* (1), (3), *and* (5), *and if*

$$u_0 \in W_0^{1,q}(\Omega), \qquad b(u_0) \in L_1^{\text{loc}}(\Omega) \cap W^{-1,q'}(\Omega), \tag{11}$$

$$f \in W^{1,1}(0, T; W^{-1,q'}(\Omega)), \tag{12}$$

the problem

$$\begin{aligned} &\frac{\partial b(u)}{\partial t} - \operatorname{div} D\Phi(\nabla u) = f \quad \text{in } \Omega \times (0, T), \\ &u = 0 \quad \text{on } \partial\Omega \times (0, T), \\ &b(u)|_{t=0} = b(u_0), \end{aligned} \tag{13}$$

admits a solution u such that

$$u \in L_\infty(0, T; W_0^{1,q}(\Omega)), \tag{14}$$

$$b(u) \in \mathscr{C}^0(0, T; L_1(\Omega)) \cap W^{1,\infty}(0, T; W^{-1,q'}(\Omega)). \tag{15}$$

Furthermore, if u_{01} *and* u_{02} *satisfy* (11) *while* f_1 *and* f_2 *satisfy* (12) *and if* $b(u_{01}) - b(u_{02})$ *is almost everywhere positive on* Ω, *while* $\langle (f_1 - f_2)(t), \varphi \rangle$ *is almost everywhere positive on* $(0, T)$ *for any* φ *in* $W_0^{1,q}(\Omega)$, *there exist a solution* u_1 *associated to* u_{01}, f_1 *and a solution* u_2 *associated to* u_{02}, f_2 *such that* $b(u_1) - b(u_2)$ *is almost everywhere positive on* $\Omega \times (0, T)$. □

Remark 2. In the settings of both theorems,

$$b(u_0)u_0 \in L_1(\Omega), \qquad b(u(t))u(t) \in L_\infty(0, T; L_1(\Omega)).$$

This property is trivially checked in the case of Theorem 1. It results from a theorem of Brézis and Browder [5, Thm. 1] in the case of Theorem 2. The positivity of Ψ and Ψ^* then implies that

$$\Psi(u_0), \Psi^*(b(u_0)) \in L_1(\Omega) \quad \text{and} \quad \Psi(u), \Psi^*(b(u)) \in L_\infty(0, T; L_1(\Omega)).$$

Furthermore in the setting of Theorem 2 the initial condition $b(u_0)$ will be shown in Remark 10 to lie in $L_1(\Omega)$, which is consistent with the continuity property of $b(u)$ with respect to time.

Remark 3. Theorem 2 provides an existence result for a class of nonlinear parabolic homogenization problems. A family A^ε of symmetric bounded measurable matrices is considered. It satisfies, for almost every x of $\mathbb{R}^N$,

$$\alpha|\xi|^2 \leqq (A^\varepsilon(x)\xi, \xi)_{\mathbb{R}^N} \leqq \beta|\xi|^2,$$

where α and β are two strictly positive real numbers.

If u_0^ε is an element of $H_0^1(\Omega)$ such that

$$b(u_0^\varepsilon) \in L_1^{\text{loc}}(\Omega) \cap H^{-1}(\Omega)$$

and if f is an arbitrary element of $W^{1,1}(0, T; H^{-1}(\Omega))$, the problem

$$\frac{\partial b(u^\varepsilon)}{\partial t} - \operatorname{div} A^\varepsilon \nabla u^\varepsilon = f \quad \text{in } \Omega,$$

$$b(u^\varepsilon)|_{t=0} = b(u_0^\varepsilon),$$

$$u^\varepsilon = 0 \quad \text{on } \partial\Omega$$

admits a solution u^ε in $L_\infty(0, T; H_0^1(\Omega))$ with $b(u^\varepsilon)$ in $L_\infty(0, T; L_1(\Omega)) \cap W^{1,\infty}(0, T; H^{-1}(\Omega))$. We assume that, as ε tends to zero,

$$u_0^\varepsilon \text{ converges weakly to } u_0^0 \text{ in } H_0^1(\Omega),$$

$$\langle b(u_0^\varepsilon), u_0^\varepsilon \rangle \text{ remains bounded independently of } \varepsilon.$$

The theory of H-convergence (Tartar [13]) ensures the existence of a subsequence $A^{\varepsilon'}$ of A^ε and of a symmetric bounded measurable matrix A^0 with

$$\alpha|\xi|^2 \leqq (A^0(x)\xi, \xi)_{\mathbb{R}^N} \leqq \beta|\xi|^2,$$

almost everywhere in $\mathbb{R}^N$, such that, as ε tends to zero,

$$A^{\varepsilon'} \text{ } H\text{-converges to } A^0.$$

It is then fairly straightforward to prove the following homogenization result: there exists a subsequence $u^{\varepsilon''}$ of u^ε such that, as ε'' tends to zero,

$$u^{\varepsilon''} \rightharpoonup u^0 \quad \text{weakly in } H_0^1(\Omega),$$

$$A^{\varepsilon''}\nabla u^{\varepsilon''} \rightharpoonup A^0 \nabla u^0 \quad \text{weakly in } [L_2(\Omega)]^N,$$

where u^0 is a solution of

$$\frac{\partial b(u^0)}{\partial t} - \operatorname{div} A^0 \nabla u^0 = f \quad \text{in } \Omega \times (0, T),$$

$$b(u^0)|_{t=0} = b(u_0^0),$$

$$u^0 = 0 \quad \text{on } \partial\Omega \times (0, T).$$

The proof of this result will not be presented here for the sake of brevity.

The proof of Theorem 1 is presented in § 2. It is divided into five steps to which correspond five sections. Section 2.1 is devoted to the formulation of a Galerkin approximation. To this effect, the function b is truncated and a small linear perturbation is added; $b(t)$ becomes $b^\eta(t) + \varepsilon t$, where b^η is the function resulting from the truncation of b at height $1/\eta$. In § 2.2 the limit process is performed in the Galerkin approximation. In § 2.3 the truncation height $1/\eta$ is increased to infinity. The coercivity parameter ε tends to zero in § 2.4. Finally the comparison result is derived in § 2.5.

The proof of Theorem 2 is presented in § 3. The initial condition u_0 is truncated at the height n while f is approximated by a sequence f^n in $W^{1,1}(0, T; L_2(\Omega))$. Theorem 1 is applied with the truncated u_0 as initial condition and with f^n as forcing term. The parameter n is then increased to infinity.

2. Proof of Theorem 1.

2.1. The Galerkin approximation. As previously mentioned, we introduce $b^\eta(t)$, $\Psi^\eta(t)$ to be

$$b^\eta(t) = \begin{cases} b(t) & \text{if } |b(t)| \leqq \dfrac{1}{\eta}, \\ \dfrac{1}{\eta}\operatorname{sg}(t) & \text{if } |b(t)| > \dfrac{1}{\eta}, \end{cases}$$

$$\Psi^\eta(t) = \int_0^t b^\eta(s)\, ds.$$

The function Ψ^η is C^1 convex and the function b^η is monotone. We propose to solve

$$\begin{aligned} &\frac{\partial}{\partial t}(b^\eta(u_\varepsilon^\eta) + \varepsilon u_\varepsilon^\eta) - \operatorname{div} D\Phi(\nabla u_\varepsilon^\eta) = f \quad \text{in } \Omega \times (0, T), \\ &u_\varepsilon^\eta = 0 \quad \text{on } \partial\Omega \times (0, T), \\ &b^\eta(u_\varepsilon^\eta)|_{t=0} = b^\eta(u_0), \end{aligned} \tag{16}$$

using a Galerkin approximation.

Let $\varphi_1, \cdots, \varphi_m, \cdots$ be a basis of $W_0^{1,q}(\Omega)$ consisting of $\mathscr{C}_0^\infty(\Omega)$-functions. If v is an arbitrary element of $W_0^{1,q}(\Omega)$, there exists a sequence V_1^m of $\mathbb{R}$ such that

$$\sum_{i=1}^m V_i^m \varphi_i \xrightarrow{m \to +\infty} v \quad \text{strongly in } W_0^{1,q}(\Omega).$$

Let m be an arbitrary but fixed integer. To any element V^m of $\mathbb{R}^m$ corresponds the element v^m of $\mathscr{C}_0^\infty(\Omega)$ defined as

$$v^m = \sum_{i=1}^m V_i^m \varphi_i.$$

The mapping is one-to-one since the φ_i are a basis of $W_0^{1,q}(\Omega)$.

Let G^m be the mapping from $\mathbb{R}^m$ into itself whose ith component is

$$[G^m(V^m)]_i = \int_\Omega (b^\eta(v^m) + \varepsilon v^m)\varphi_i\, dx.$$

If V^m and W^m are two arbitrary elements of $\mathbb{R}^m$,

$$\begin{aligned} (G^m(V^m) - G^m(W^m),\, V^m - W^m)_{\mathbb{R}^m} &= \int_\Omega (b^\eta(v^m) - b^\eta(w^m))(v^m - w^m)\, dx \\ &\quad + \varepsilon \int_\Omega |v^m - w^m|^2\, dx \geqq \varepsilon \mathscr{C}_m |V^m - W^m|_{\mathbb{R}^m}^2, \end{aligned} \tag{17}$$

where $\mathscr{C}_m$ is a constant such that for any V^m in $\mathbb{R}^m$

$$\mathscr{C}_m |V^m|_{\mathbb{R}^m}^2 \leqq \int_\Omega |v^m|^2\, dx.$$

Hence G^m is monotone coercive; it is also clearly continuous. We thus conclude with the help of Brouwer's fixed point theorem that G^m is onto (cf., for example, [8, Lemma 4.3, p. 53]). In view of (17),

$$(G^m)^{-1} \text{ is Lipschitz with Lipschitz ratio } 1/\varepsilon. \tag{18}$$

Let Φ^m be the mapping from $\mathbb{R}^m$ into itself whose ith component is

$$[\Phi^m(V^m)]_i = \int_\Omega (D\Phi(\nabla v^m), \nabla \varphi_i)_{\mathbb{R}^N}\, dx.$$

Since $D\Phi$ is continuous on $[L_q(\Omega)]^m$, Φ^m is continuous.

Finally define $F^m(t)$ to be the vector of $\mathbb{R}^m$ whose ith component is

$$[F^m(t)]_i = \int_\Omega f(t)\varphi_i\, dx.$$

By virtue of (7), $F^m(t)$ is a continuous function of t.

The continuity properties of $(G^m)^{-1}$, Φ^m, and F^m imply that the ordinary differential equation

$$\frac{dW^m}{dt}(t) + \Phi^m((G^m)^{-1}(W^m))(t) = F^m(t), \qquad W^m(0) = W_0^m, \tag{19}$$

where W_0^m is an arbitrary element of $\mathbb{R}^m$, has a local C^1 solution $W^m(t)$ on a time interval $[0, T(W_0^m))$; $T(W_0^m)$ is a strictly positive real number which depends on W_0^m.

The existence of a global solution on $[0, T]$ is ensured if $|W^m(t)|$ is proved not to blow up whenever t tends to $T(W_0^m)$ with $T(W_0^m) \leqq T$. In view of the continuity properties of G^m, it suffices to show that $V^m(t)$, defined as

$$V^m(t) = (G^m)^{-1}(W^m(t)),$$

has a bounded norm as t tends to $T(W_0^m)$.

If we set

$$V_0^m = (G^m)^{-1}(W_0^m),$$

the system (19) reads as follows:

$$\frac{d}{dt} G^m(V^m(t)) + \Phi^m(V^m(t)) = F^m(t), \qquad V^m(0) = V_0^m. \tag{20}$$

Multiplication of the first equality of (20) by $V^m(t)$ and integration over the time interval $(0, t)$ of the resulting expression leads to

$$\int_0^t \int_\Omega \left(\frac{\partial b^\eta(v^m)}{\partial t} v^m + \varepsilon \frac{\partial v^m}{\partial t} v^m \right)(s)\, dx\, ds + \int_0^t \int_\Omega (D\Phi(\nabla v^m(s)), \nabla v^m(s))_{\mathbb{R}^N}\, dx\, ds$$
$$= \int_0^t \int_\Omega f(s) v^m(s)\, dx\, ds. \tag{21}$$

We now appeal to the following result first established by Alt and Luckhaus [1, Lemma 1.5, p. 315].

LEMMA 1. *Let Ω be a bounded domain of $\mathbb{R}^n$. Let $\tilde{\Psi}$ be a C^1 convex function on $\mathbb{R}$, with $\tilde{b}$ as derivative ($\tilde{\Psi}(0) = 0$). Let $\tilde{\Psi}^*$ be its convex conjugate. Assume that*

$$\begin{aligned} &u \in L_\infty(0, T; W_0^{1,s}(\Omega)), \qquad 1 < s < +\infty, \\ &\tilde{b}(u) \in L_\infty(0, T; L_1(\Omega)), \\ &\frac{\partial}{\partial t}\tilde{b}(u) \in L_\infty(0, T; W^{-1,s'}(\Omega)), \qquad \frac{1}{s} + \frac{1}{s'} = 1. \end{aligned} \tag{22}$$

Assume further that there exists an element u_0 *in* $W_0^{1,s}(\Omega)$ *such that*

$$\tilde{b}(u)|_{t=0} = \tilde{b}(u_0), \tag{23}$$

and that

$$\tilde{b}(u_0) \in L_1^{\mathrm{loc}}(\Omega) \cap W^{-1,s'}(\Omega). \tag{24}$$

Then

$$\tilde{\Psi}^*(\tilde{b}(u)) \in L^\infty(0, T; L_1(\Omega)), \qquad \tilde{\Psi}^*(\tilde{b}(u_0)) \in L_1(\Omega) \tag{25}$$

and, for almost any t *in* $(0, T)$,

$$\int_\Omega \tilde{\Psi}^*(b(u(t)))\, dx - \int_\Omega \tilde{\Psi}^*(\tilde{b}(u_0))\, dx = \int_0^t \left\langle \frac{\partial \tilde{b}(u(s))}{\partial t}, u(s) \right\rangle ds, \tag{26}$$

where $\langle\, , \rangle$ *stands for the duality bracket between* $W_0^{1,s}(\Omega)$ *and* $W^{-1,s'}(\Omega)$. □

Lemma 1 is applied in our context with $s = q$, $\tilde{\Psi}(t) = \Psi^\eta(t) + \varepsilon(t^2/2)$, $T = t < T(W_0^m)$, $u_0 = v_0^m$ and it yields

$$\begin{aligned} &\int_0^t \int_\Omega \frac{\partial}{\partial t}(b^\eta(v^m(s)) + \varepsilon v^m(s)) v^m(s)\, dx\, ds \\ &\quad = \int_\Omega (\Psi^\eta)^*(b^\eta(v^m(t)))\, dx + \frac{\varepsilon}{2}\|v^m(t)\|_{0,2}^2 - \frac{\varepsilon}{2}\|v_0^m\|_0^2 - \int_\Omega (\psi^\eta)^*(b^\eta(v_0^m))\, dx. \end{aligned} \tag{27}$$

Remark 4. In view of the regularity of the function v^m relation (27) can be established independently of Lemma 1. At a later stage of this study however, Lemma 1 will become an essential ingredient and it will be repeatedly applied with $s = q$.

Inserting (27) into (21) leads to

$$\begin{aligned} &\int_\Omega (\Psi^\eta)^*(b^\eta(v^m(t)))\, dx + \frac{\varepsilon}{2}\|v^m(t)\|_{0,2}^2 + \int_0^t \int_\Omega (D\Phi(\nabla v^m(s)), \nabla v^m(s))_{\mathbb{R}^N}\, dx\, ds \\ &\quad \leqq \int_\Omega (\Psi^\eta)^*(b^\eta(v_0^m))\, dx + \frac{\varepsilon}{2}\|v_0^m\|_{0,2}^2 + |||f||| \int_0^t \|v^m(s)\|_{1,q}\, ds, \end{aligned} \tag{28}$$

where from now on $|||f|||$ stands for the norm of f in $W^{1,1}(0, T; W^{-1,q'}(\Omega))$.

Since

$$\Phi(0) = 0,$$

the coercivity property of Φ (cf. (5)) together with Poincaré's inequality imply that

$$\int_0^t \int_\Omega (D\Phi(\nabla v^m(s)), \nabla v^m(s))_{\mathbb{R}^N}\, dx\, ds \geqq \int_0^t \Phi(\nabla v^m(s))\, ds \geqq \alpha \int_0^t \|v^m(s)\|_{1,q}^r\, ds. \tag{29}$$

Because $(\Psi^\eta)^*$ is always positive, insertion of (25) and (29) into (28) yields

$$\frac{\varepsilon}{2}\|v^m(t)\|_{0,2}^2 + \alpha \int_0^t \|v^m(s)\|_{1,q}^r\, ds \leqq \mathscr{C} + |||f||| \int_0^t \|v^m(s)\|_{1,q}\, ds, \tag{30}$$

where $\mathscr{C}$ denotes a generic constant.

Since r is strictly greater than one, (30) implies that $\|v^m(t)\|_{0,2}$ remains bounded on $[0, T(W_0^m))$ and thus that $|V^m(t)|$ remains bounded as t tends to $T(W_0^m)$, which was the result sought.

Recalling (6), we denote by U_0^m the vector of $\mathbb{R}^m$ associated with the projection u_0^m of u_0 on the span of $\varphi_1, \cdots, \varphi_m$, i.e.,

$$u_0^m = \sum_{i=1}^{m} U_{01}^m \varphi_i \xrightarrow{m \to +\infty} u_0 \quad \text{strongly in } W_0^{1,q}(\Omega).$$

According to the previous analysis, the equation

$$\frac{d}{dt} G^m(U^m(t)) + \Phi^m(U^m(t)) = F^m(t), \qquad U^m(0) = U_0^m, \tag{31}$$

admits a global Lipschitz solution on $[0, T]$.

A priori estimates. Multiplication of the first equality of (31) by dU^m/dt and integration over the time interval $(0, t)$ of the resulting expression leads to

$$\begin{aligned} &\int_0^t \int_\Omega \frac{\partial b^\eta(u^m(s))}{\partial t} \frac{\partial u^m(s)}{\partial t}\, dx\, ds + \varepsilon \int_0^t \int_\Omega \left|\frac{\partial u^m(s)}{\partial t}\right|^2 dx\, ds + \Phi(\nabla u^m(t)) \\ &\quad = \Phi(\nabla u_0^m) + \int_0^t \int_\Omega f(s) \frac{\partial u^m(s)}{\partial t}\, dx\, ds, \end{aligned} \tag{32}$$

where

$$u^m(t) = \sum_{i=1}^{m} U_i^m(t) \varphi_i.$$

The function b^η is Lipschitz and u^m is in $\mathrm{Lip}\,(\Omega \times (0, T))$; thus, by virtue of the monotonicity of b^η,

$$\int_0^t \int_\Omega \frac{\partial b^\eta(u^m(s))}{\partial t} \frac{\partial u^m(s)}{\partial t}\, dx\, ds = \int_0^t \int_\Omega (b^\eta)'(u^m(s)) \left(\frac{\partial u^m(s)}{\partial t}\right)^2 dx\, ds \geqq 0. \tag{33}$$

The coercivity property of Φ (cf. (5)), (in)equalities (32), (33), and Poincaré's inequality imply that

$$\int_0^t \left\| \sqrt{\varepsilon} \frac{\partial u^m}{\partial t} \right\|_{0,2}^2 ds + \alpha \|u^m(t)\|_{1,q}^r \leqq \Phi(\nabla u_0^m) + \int_0^t \int_\Omega f(s) \frac{\partial u^m(s)}{\partial t}\, dx\, ds. \tag{34}$$

The last term of the right-hand side of inequality (34) is integrated by parts with the help of (7). We obtain

$$\int_0^t \left\| \sqrt{\varepsilon} \frac{\partial u^m}{\partial t} \right\|_{0,2}^2 ds + \alpha \|u^m(t)\|_{1,q}^r \leqq \Phi(\nabla u_0^m) + 3 |||f||| \sup_{s \in [0,t]} \|u^m(s)\|_{1,q}. \tag{35}$$

The time t can be chosen arbitrarily in $[0, T]$; thus,

$$\begin{aligned} &\int_0^T \left\| \sqrt{\varepsilon} \frac{\partial u^m(s)}{\partial t} \right\|_{0,2}^2 ds + \alpha \sup_{t \in [0,T]} \|u^m(t)\|_{1,q}^r \\ &\quad \leqq \Phi(\nabla u_0^m) + 3 |||f||| \sup_{t \in [0,T]} \|u^m(t)\|_{1,q}. \end{aligned} \tag{36}$$

Finally ∇u_0^m converges to ∇u_0 in $[L_q(\Omega)]^N$ as m goes to infinity. The continuity of Φ on $[L_q(\Omega)]^N$ implies that

$$\Phi(\nabla u_0^m) \xrightarrow{m \to +\infty} \Phi(\nabla u^0).$$

If m is taken to be large enough, (36) reads as follows:

$$\int_0^T \left\| \sqrt{\varepsilon} \frac{\partial u^m}{\partial t} \right\|_{0,2}^2 ds + \alpha \sup_{t\in[0,T]} \|u^m(t)\|_{1,q}^r \leqq \Phi(\nabla u_0) + 1 + 3|||f||| \sup_{t\in[0,T]} \|u^m(t)\|_{1,q}. \tag{37}$$

The function $\|u^m(t)\|_{1,q}$ is continuous on $[0, T]$; it reaches its supremum on $[0, T]$. It is then easily deduced from (37) that

(38) $\sqrt{\varepsilon}\, \partial u^m/\partial t$ is bounded in $L_2(0, T; L_2(\Omega))$ independently of m, η, or ε,

(39) $u^m(t)$ is bounded in $L_\infty(0, T; W_0^{1,q}(\Omega))$ independently of m, η, or ε.

Because $D\Phi$ is bounded on the bounded sets of $[L_q(\Omega)]^N$, (39) implies that, if $1/q' + 1/q = 1$,

(40) $D\Phi(\nabla u^m)$ is bounded in $L_\infty(0, T; [L_{q'}, (\Omega)]^N)$ independently of m, η, or ε.

Remark 5. The bounds in (38)–(40) continuously depend on $\Phi(\nabla u_0)$ and $|||f|||$ only (recall that $|||f|||$ is the $W^{1,1}(0, T; W^{-1,q'}(\Omega))$-norm of f).

Finally, since b^η is bounded on $\mathbb{R}$,

(41) $b^\eta(u^m)$ is bounded in $L_\infty((0, T)\times\Omega)$,

but the bound depends on η.

With the help of (38)–(41) we conclude that there exists a suitably extracted subsequence of u^m (still denoted u^m) such that, as m tends to infinity,

$$\begin{aligned} &u^m \rightharpoonup u_\varepsilon^\eta \quad \text{weak-* in } L_\infty(0, T; W_0^{1,q}(\Omega)), \\ &\sqrt{\varepsilon}\frac{\partial u^m}{\partial t} \rightharpoonup \sqrt{\varepsilon}\frac{\partial u_\varepsilon^\eta}{\partial t} \quad \text{weakly in } L_2(0, T; L_2(\Omega)), \\ &D\Phi(\nabla u^m) \rightharpoonup Y_\varepsilon^\eta \quad \text{weak-* in } L_\infty(0, T; [L_{q'}(\Omega)]^N), \\ &b^\eta(u^m) \rightharpoonup \chi_\varepsilon^\eta \quad \text{weak-* in } L_\infty((0, T)\times\Omega). \end{aligned} \tag{42}$$

In the following section we propose to use (42) to pass to the limit in (31) as m tends to infinity, and to identify the quantities Y_ε^η and χ_ε^η.

2.2. Passing to the limit in the Galerkin approximation. Let ζ be an arbitrary element of $\mathscr{C}_0^\infty((0, T))$. Equation (31) together with the convergence (42) imply that, for any integer i,

$$\left\langle\!\!\left\langle \frac{\partial \chi_\varepsilon^\eta}{\partial t} + \varepsilon \frac{\partial u_\varepsilon^\eta}{\partial t} - \operatorname{div} Y_\varepsilon^\eta - f, \varphi_i \zeta \right\rangle\!\!\right\rangle = 0,$$

where the duality bracket refers to the duality between $\mathscr{C}_0^\infty((0, T)\times\Omega)$ and $\mathscr{D}'((0, T)\times\Omega)$ (the basis vectors φ_i lie in $\mathscr{C}_0^\infty(\Omega)$).

The sequence $\{\varphi_i\}$ is a basis of $W_0^{1,q}(\Omega)$ which contains $\mathscr{C}_0^\infty(\Omega)$. Thus, if φ is an arbitrary element of $\mathscr{C}_0^\infty(\Omega)$,

$$\left\langle\!\!\left\langle \frac{\partial \chi_\varepsilon^\eta}{\partial t} + \varepsilon \frac{\partial u_\varepsilon^\eta}{\partial t} - \operatorname{div} Y_\varepsilon^\eta - f, \varphi \zeta \right\rangle\!\!\right\rangle = 0,$$

and, by the density of $\mathscr{C}_0^\infty((0, T))\times\mathscr{C}_0^\infty(\Omega)$ in $\mathscr{C}_0^\infty((0, T)\times\Omega)$,

$$\frac{\partial \chi_\varepsilon^\eta}{\partial t} + \varepsilon \frac{\partial u_\varepsilon^\eta}{\partial t} - \operatorname{div} Y_\varepsilon^\eta - f = 0. \tag{43}$$

In view of (43), χ_ε^η has a trace in $W^{-1,q'}(\Omega)$. A proper choice of ζ in $\mathscr{C}_0^\infty([0, T))$ and convergence (42) lead to

$$\langle\langle \chi_\varepsilon^\eta(0) + \varepsilon u_\varepsilon^\eta(0)\varphi_i\zeta(0)\rangle\rangle = \lim_{m\to+\infty} \int_\Omega (b^\eta(u_0^m) + \varepsilon u_0^m)\varphi_i\zeta(0).$$

But $\{\varphi_i\}$ is a basis of $W_0^{1,q}(\Omega)$ and u_0^m converges strongly to u_0 in $W_0^{1,q}(\Omega)$; thus

$$\chi_\varepsilon^\eta(0) + \varepsilon u_\varepsilon^\eta(0) = b^\eta(u_0) + \varepsilon u_0. \tag{44}$$

Because of the estimates on u^m and $\partial u^m/\partial t$ (cf. (42))

$$u^m(0) \rightharpoonup u_\varepsilon^\eta(0) \quad \text{weakly in } L_2(\Omega)$$

as m tends to infinity; thus

$$u_\varepsilon^\eta(0) = u_0, \tag{45}$$

and (44) yields

$$\chi_\varepsilon^\eta(0) = b^\eta(u_0). \tag{46}$$

We now seek to identify the quantities χ_ε^η and Y_ε^η. The identification of χ_ε^η is very simple. A straightforward application of Aubin's lemma (cf., e.g., [12, Cor. 4]) to u^m implies that

$$u^m \to u_\varepsilon^\eta \quad \text{strongly in } \mathscr{C}^0([0, T]; L_2(\Omega))$$

as m tends to infinity. Since b^η is Lipschitz and bounded, it follows that for any finite s

$$b^\eta(u^m) \to b^\eta(u_\varepsilon^\eta) \quad \text{strongly in } \mathscr{C}^0([0, T]; L_s(\Omega))$$

as m tends to infinity and, in view of (42), that

$$\chi_\varepsilon^\eta = b^\eta(u_\varepsilon^\eta), \qquad b^\eta(u_\varepsilon^\eta) \in \mathscr{C}^0([0, T]; L_s(\Omega)), \qquad 1 \leqq s < +\infty. \tag{47}$$

The identification of Y_ε^η is performed with the help of the following simple lemma.

LEMMA 2. *Assume that* Φ *satisfies* (5) *and that* w_m *is a sequence of* $L_\infty(0, T; [L_q(\Omega)]^N)$ *such that*

$$\begin{aligned} &w_m \rightharpoonup w \quad \textit{weak-* in } L_\infty(0, T; [L_q(\Omega)]^N), \\ &D\Phi(w_m) \rightharpoonup Y \quad \textit{weak-* in } L_\infty(0, T; [L_{q'}(\Omega)]^N), \end{aligned} \tag{48}$$

as m *tends to infinity (or zero). If*

$$\int_0^T \int_0^t \int_\Omega (Y(s), w(s))_{\mathbb{R}^N}\, dx\, ds\, dt \geqq \overline{\lim} \int_0^T \int_0^t \int_\Omega (D\Phi(w_m(s), w_m(s)))\, dx\, ds\, dt, \tag{49}$$

then

$$Y = D\Phi(w). \tag{50}$$

The proof of this lemma is a straightforward adaptation of a classical result of Minty (cf., e.g., [8, Remark 2.1, p. 173 and Prop. 2.5, p. 179]). It will not be reproduced here.

In our setting w_m, w, and Y are to be identified with ∇u^m, $\nabla u_\varepsilon^\eta$, and Y_ε^η, respectively. In view of (42), (48) is satisfied. To show that

$$Y_\varepsilon^\eta = D\Phi(\nabla u_\varepsilon^\eta) \tag{51}$$

we only need to prove that

$$\int_0^T \int_0^t \int_\Omega (Y_\varepsilon^\eta(s), \nabla u_\varepsilon^\eta(s))_{\mathbb{R}^N} \, dx \, ds \, dt \tag{52}$$
$$\geqq \overline{\lim_{m \to +\infty}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u^m(s)), \nabla u^m(s))_{\mathbb{R}^N} \, dx \, ds \, dt.$$

According to (21) and (27) (applied to u^m in place of v^m), we have

$$\int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u^m(s)), \nabla u^m(s))_{\mathbb{R}^N} \, dx \, ds \, dt$$
$$= \int_0^T \int_0^t \int_\Omega f(s) u^m(s) \, dx \, ds \, dt + T \left[\int_\Omega (\Psi^\eta)^*(b^\eta(u_0^m)) \, dx + \frac{\varepsilon}{2} \|u_0^m\|_{0,2}^2 \right] \tag{53}$$
$$- \int_0^T \left[\int_\Omega (\Psi^\eta)^*(b^\eta(u^m(t))) \, dx + \frac{\varepsilon}{2} \|u^m(t)\|_{0,2}^2 \right] dt.$$

Note that Ψ^η is Lipschitz and that (4) holds for b^η, Ψ^η, and $(\Psi^\eta)^*$ in place of b, Ψ, and Ψ^*, respectively. Since u^m converges weak-* to u_ε^η in $L_\infty(0, T; W_0^{1,q}(\Omega))$ and u_0^m converges strongly in $W_0^{1,q}(\Omega)$ to u_0, the two first terms of the right-hand side of equality (53) are easily seen to converge to

$$\int_0^T \int_0^t \int_\Omega f(s) u_\varepsilon^\eta(s) \, dx \, ds \, dt + T \left[\int_\Omega (\Psi^\eta)^*(b^\eta(u_0)) \, dx + \frac{\varepsilon}{2} \|u_0\|_{0,2}^2 \right]$$

as m goes to infinity.

The strong convergences of the sequences $b^\eta(u^m)$ and u^m in $\mathscr{C}^0([0, T]; L_2(\Omega))$ imply that

$$\lim_{m \to +\infty} - \int_0^T \int_\Omega (\Psi^\eta)^*(b^\eta(u^m(t))) \, dx \, dt = - \int_0^T \int_\Omega (\Psi^\eta)^*(b^\eta(u_\varepsilon^\eta(t))) \, dx \, dt$$

and

$$\lim_{m \to +\infty} \int_0^T \|u^m(t)\|_{0,2}^2 \, dt = \int_0^T \|u_\varepsilon^\eta(t)\|_{0,2}^2 \, dt.$$

We are now in a position to pass to the limit in (53). We obtain the following:

$$\lim_{m \to +\infty} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u^m(s)), \nabla u^m(s))_{\mathbb{R}^N} \, dx \, ds \, dt$$
$$= \int_0^T \int_0^t \int_\Omega f(s) u_\varepsilon^\eta(s) \, dx \, ds \, dt + T \left[\int_\Omega (\Psi^\eta)^*(b^\eta(u_0)) \, dx + \frac{\varepsilon}{2} \|u_0\|_{0,2}^2 \right] \tag{54}$$
$$- \int_0^T \left[\int_\Omega (\Psi^\eta)^*(b^\eta(u_\varepsilon^\eta(t))) \, dx + \frac{\varepsilon}{2} \|u_\varepsilon^\eta(t)\|_{0,2}^2 \right] dt.$$

Multiplication of (43) by u_ε^η and integration of the resulting expression over $(0, t) \times \Omega$, then over $(0, T)$ yields, with the help of (47),

$$\int_0^T \int_0^t \int_\Omega (Y_\varepsilon^\eta(s), \nabla u_\varepsilon^\eta(s))_{\mathbb{R}^N} \, dx \, ds \, dt$$
$$= \int_0^T \int_0^t \int_\Omega f(s) u_\varepsilon^\eta(s) \, dx \, ds \, dt \tag{55}$$
$$- \int_0^T \int_0^t \left\langle \frac{\partial}{\partial t} (b^\eta(u_\varepsilon^\eta(s)) + \varepsilon u_\varepsilon^\eta(s)), u_\varepsilon^\eta(s) \right\rangle ds \, dt.$$

The last term in the right-hand side of equality (55) is evaluated with the help of Lemma 1 (cf. Remark 4). We obtain the following:

$$\begin{aligned}&\int_0^T \int_0^t \left\langle \frac{\partial}{\partial t}(b^\eta(u_\varepsilon^\eta(s)) + \varepsilon u_\varepsilon^\eta(s)),\, u_\varepsilon^\eta(s) \right\rangle ds\, dt \\ &\quad = \int_0^T \int_\Omega (\Psi^\eta)^*(b^\eta(u_\varepsilon^\eta(t)))\, dx\, dt \\ &\quad\quad + \frac{\varepsilon}{2}\int_0^T \|u_\varepsilon^\eta(t)\|_{0,2}^2\, dt - T\frac{\varepsilon}{2}\|u_0\|_{0,2}^2 - T\int_\Omega (\Psi^\eta)^*(b^\eta(u_0))\, dx.\end{aligned} \tag{56}$$

Inserting (56) into (55) and comparing the resulting expression with the right-hand side of inequality (54) yields (52), which in turn proves (51).

At this stage of the proof we have constructed a sequence u_ε^η with the following properties:

$$u_\varepsilon^\eta \in L_\infty(0, T; W_0^{1,q}(\Omega)), \qquad b^\eta(u_\varepsilon^\eta) \in L_\infty((0, T)\times\Omega),$$

$$\frac{\partial b^\eta(u_\varepsilon^\eta)}{\partial t} + \varepsilon\frac{\partial u_\varepsilon^\eta}{\partial t} - \operatorname{div} D\Phi(\nabla u_\varepsilon^\eta) = f, \tag{57}$$

$$u_\varepsilon^\eta(0) = u_0, \tag{58}$$

$$b^\eta(u_\varepsilon^\eta)|_{t=0} = b^\eta(u_0). \tag{59}$$

Furthermore the weak lower semicontinuity properties of the L_2, $L_{q'}$, and L_q norms imply, by virtue of (38)–(40) and (42), that

(60) $\sqrt{\varepsilon}\,\partial u_\varepsilon^\eta/\partial t$ is bounded in $L_2(0, T; L_2(\Omega))$ independently of ε and η,

(61) u_ε^η is bounded in $L_\infty(0, T; W_0^{1,q}(\Omega))$ independently of ε and η,

(62) $D\Phi(\nabla u_\varepsilon^\eta)$ is bounded in $L_\infty(0, T; [L_{q'}(\Omega)]^N)$ independently of ε and η.

With the help of (60)–(62) we conclude that there exists a suitably extracted subsequence (still denoted u_ε^η) such that, for fixed ε,

$$\begin{aligned}&u_\varepsilon^\eta \rightharpoonup u_\varepsilon \quad \text{weak-* in } L_\infty(0, T; W_0^{1,q}(\Omega)), \\ &\sqrt{\varepsilon}\frac{\partial u_\varepsilon^\eta}{\partial t} \rightharpoonup \sqrt{\varepsilon}\frac{\partial u_\varepsilon}{\partial t} \quad \text{weakly in } L_2(0, T; L_2(\Omega)), \\ &D\Phi(\nabla u_\varepsilon^\eta) \rightharpoonup Y_\varepsilon \quad \text{weak-* in } L_\infty(0, T; [L_{q'}(\Omega)]^N),\end{aligned} \tag{63}$$

as η goes to zero.

In the following section we propose to use (63) to pass to the limit in (57)–(59) as η tends to zero, and to identify the quantity Y_ε. To this effect we need to derive an estimate on $b^\eta(u_\varepsilon^\eta)$ independent of η (and ε) and to identify its weak limit.

2.3. Passing to the limit in the truncation. The quantities $\Psi^\eta(u_\varepsilon^\eta)$, $b^\eta(u_\varepsilon^\eta)$ lie in $L_\infty(0, T; W_0^{1,q}(\Omega)) \cap W^{1,2}(0, T; L_2(\Omega))$ because Ψ^η and b^η are Lipschitz. Thus, $b^\eta(u_\varepsilon^\eta)$ is an admissible test function in (57). Upon integration over the time interval $(0, t)$ of the result of the multiplication of (57) by $b^\eta(u_\varepsilon^\eta)$, we obtain

$$\begin{aligned}&\frac{1}{2}\|b^\eta(u_\varepsilon^\eta(t))\|_{0,2}^2 + \varepsilon\int_\Omega \Psi^\eta(u_\varepsilon^\eta(t))\, dx \\ &\quad + \int_0^t \int_\Omega (b^\eta)'(u_\varepsilon^\eta(s))(D\Phi(\nabla u_\varepsilon^\eta(s)), \nabla u_\varepsilon^\eta(s))_{\mathbb{R}^N}\, dx\, ds \\ &\quad = \frac{1}{2}\|b^\eta(u_0)\|_{0,2}^2 + \varepsilon\int_\Omega \Psi^\eta(u_0)\, dx + \int_0^t \int_\Omega f(s) b^\eta(u_\varepsilon^\eta(s))\, dx\, ds.\end{aligned} \tag{64}$$

The derivation formula for the composition of a $W^{1,q}$ function by a Lipschitz function is implicitly used in (64). It is classical if the Lipschitz function is piecewise C^1. For a proof in the case of an arbitrary Lipschitz function see, for example, [9, Cor. 1.3, p. 353] or [4, Thm. 4.3].

Since b^η is monotone and Ψ^η is positive, (64) yields

$$\begin{aligned}\frac{1}{2}\|b^\eta(u_\varepsilon^\eta(t))\|_{0,2}^2 \leqq \frac{1}{2}\|b^\eta(u_0)\|_{0,2}^2 + \varepsilon \int_\Omega \Psi^\eta(u_0)\,dx \\ + T \sup_{t\in[0,T]} \|f(t)\|_{0,2} \sup_{t\in[0,T]} \|b^\eta(u_\varepsilon^\eta(t))\|_{0,2}.\end{aligned} \tag{65}$$

Remark 6. The L_2 space regularity of f is required in estimate (65), and it motivates the regularity hypothesis (7) on f.

As η tends to zero, $b^\eta(u_0)$ and $\Psi^\eta(u_0)$ converge almost everywhere to $b(u_0)$ and $\Psi(u_0)$, respectively, while

$$|b^\eta(u_0(x))| \leqq |b(u_0(x))|, \qquad \Psi^\eta(u_0(x)) \leqq \Psi(u_0(x)),$$

for almost every x of Ω. By hypothesis, $b(u_0)$ belongs to $L_2(\Omega)$ (see (6)). In view of (4) and the positivity of Ψ,

$$0 \leqq \Psi(u_0(x)) \leqq b(u_0(x))u_0(x)$$

for almost every x of Ω, and thus $\Psi(u_0)$ belongs to $L_1(\Omega)$.

The dominated convergence theorem permits us to conclude that

$$\begin{aligned} b^\eta(u_0) &\to b(u_0) \qquad \text{strongly in } L_2(\Omega), \\ \Psi^\eta(u_0) &\to \Psi(u_0) \qquad \text{strongly in } L_1(\Omega), \end{aligned} \tag{66}$$

and thus to give an upper bound for the right-hand side of (65). We obtain, for η small enough,

$$\begin{aligned}\frac{1}{2}\|b^\eta(u_\varepsilon^\eta(t))\|_{0,2}^2 \leqq \frac{1}{2}\|b(u_0)\|_{0,2}^2 + \varepsilon \int_\Omega \Psi(u_0)\,dx + 1 \\ + T \sup_{t\in[0,T]} \|f(t)\|_{0,2} \sup_{t\in[0,T]} \|b^\eta(u_\varepsilon^\eta(t))\|_{0,2}.\end{aligned} \tag{67}$$

Since $\|b^\eta(u_\varepsilon^\eta(t))\|_{0,2}$ is continuous on $[0, T]$ (cf. (47)), an argument similar to the one that led to (38), (39), would show that

$$b^\eta(u_\varepsilon^\eta) \text{ is bounded in } L_\infty(0, T, L_2(\Omega)) \text{ independently of } \varepsilon \text{ and } \eta. \tag{68}$$

At the possible expense of extracting one more subsequence, we are thus at liberty to assume that the sequence u_ε^η is also such that

$$b^\eta(u_\varepsilon^\eta) \rightharpoonup \chi_\varepsilon \quad \text{weak-*} \text{ in } L_\infty(0, T; L_2(\Omega)) \tag{69}$$

as η tends to zero.

Passing to the limit in (57)-(59) is now an immediate task if we remark that, in view of (57), (60), and (62),

$$\partial b^\eta(u_\varepsilon^\eta)/\partial t \text{ is bounded in } L_2(0, T; W^{-1,q'}(\Omega)). \tag{70}$$

We obtain the following:

$$\frac{\partial \chi_\varepsilon}{\partial t} + \varepsilon \frac{\partial u_\varepsilon}{\partial t} - \operatorname{div} Y_\varepsilon = f, \tag{71}$$

and, with the help of (60), (61), (68), and (70),

$$u_\varepsilon(0) = u_0, \tag{72}$$

$$\chi_\varepsilon|_{t=0} = b(u_0). \tag{73}$$

It now remains to identify χ_ε and Y_ε. Once again, the identification of χ_ε directly results from Aubin's lemma. In view of (63),

$$u_\varepsilon^\eta \to u_\varepsilon \quad \text{strongly in } \mathscr{C}^0([0, T]; L_2(\Omega))$$

as η tends to zero. Since b^η converges pointwise to b on $\mathbb{R}$ as η tends to zero, a subsequence of $b^\eta(u_\varepsilon^\eta)$ (still denoted $b^\eta(u_\varepsilon^\eta)$) satisfies

$$b^\eta(u_\varepsilon^\eta(x, t)) \to b(u_\varepsilon(x, t)) \text{ for almost every } (x, t) \text{ of } \Omega \times (0, T)$$

as η tends to zero. Recalling (69) then implies that

$$\chi_\varepsilon = b(u_\varepsilon), \qquad b(u_\varepsilon) \in L_\infty(0, T; L_2(\Omega)). \tag{74}$$

The identification of Y_ε relies on Lemma 2. The quantities w_m, w, and Y are identified with $\nabla u_\varepsilon^\eta$, ∇u_ε, and Y_ε, respectively. In view of (63), (48) is satisfied. To show that

$$Y_\varepsilon = D\Phi(\nabla u_\varepsilon), \tag{75}$$

we only need to prove that

$$\begin{aligned} &\int_0^T \int_0^t \int_\Omega (Y_\varepsilon(s), \nabla u_\varepsilon(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\qquad \geqq \overline{\lim_{\eta \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u_\varepsilon^\eta(s)), \nabla u_\varepsilon^\eta(s))_{\mathbb{R}^N} \, dx \, ds \, dt. \end{aligned} \tag{76}$$

The proof of (76) is essentially the same as that of (52). It consists in passing to the limit in (55), (56) as η tends to zero (with Y_ε^η replaced by $D\Phi(\nabla u_\varepsilon^\eta)$). We obtain the following:

$$\begin{aligned} &\overline{\lim_{\eta \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u_\varepsilon^\eta(s)), \nabla u_\varepsilon^\eta(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\qquad \leqq \int_0^T \int_0^t \int_\Omega f(s) u_\varepsilon(s) \, dx \, ds \, dt + T\frac{\varepsilon}{2} \|u_0\|_{0,2}^2 - \frac{\varepsilon}{2} \int_0^T \|u_\varepsilon(t)\|_{0,2}^2 \, dt \\ &\qquad - \underline{\lim_{\eta \to 0}} \int_0^T \int_\Omega (\Psi^\eta)^*(b^\eta(u_\varepsilon^\eta(t))) \, dx \, dt + T \lim_{\eta \to 0} \int_\Omega (\Psi^\eta)^*(b^\eta(u_0)) \, dx. \end{aligned} \tag{77}$$

Since

$$(\Psi^\eta)^*(b^\eta(u_0)) = u_0 b^\eta(u_0) - \Psi^\eta(u_0),$$

convergences (66) imply that

$$\lim_{\eta \to 0} \int_\Omega (\Psi^\eta)^*(b^\eta(u_0)) \, dx = \int_\Omega (u_0 b(u_0) - \Psi(u_0)) \, dx = \int_\Omega \Psi^*(b(u_0)) \, dx. \tag{78}$$

Similarly, since u_ε^η and $b^\eta(u_\varepsilon^\eta)$ converge almost everywhere to u_ε and $b(u_\varepsilon)$, respectively,

$$(\Psi^\eta)^*(b^\eta(u_\varepsilon^\eta(x, t))) \to \Psi^*(b(u_\varepsilon(x, t)))$$

for almost any (x, t) in $\Omega \times (0, T)$ as η tends to zero. The positivity of $(\Psi^\eta)^*$ together with Fatou's lemma yield

$$\varliminf_{\eta \to 0} \int_0^T \int_\Omega (\Psi^\eta)^*(b^\eta(u_\varepsilon^\eta(t)))\, dx\, dt \geqq \int_0^T \int_\Omega \Psi^*(b(u_\varepsilon(t)))\, dx\, dt. \tag{79}$$

Insertion of (78) and (79) into (77) leads to

$$\begin{aligned} &\varlimsup_{\eta \to 0} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u_\varepsilon^\eta(s)), \nabla u_\varepsilon^\eta(s))_{\mathbb{R}^N}\, dx\, ds\, dt \\ &\leqq \int_0^T \int_0^t \int_\Omega f(s) u_\varepsilon(s)\, dx\, ds\, dt + T\left[\int_\Omega \Psi^*(b(u_0))\, dx + \frac{\varepsilon}{2}\|u_0\|_{0,2}^2\right] \\ &\quad - \int_0^T \left[\int_\Omega \Psi^\eta(b(u_\varepsilon(t)))\, dx + \frac{\varepsilon}{2}\|u_\varepsilon(t)\|_{0,2}^2\right] dt. \end{aligned} \tag{80}$$

Multiplication of (71) by u_ε, integration of the resulting expression over $(0, t) \times \Omega$ then over $(0, T)$, and application of Lemma 1 readily shows that the right-hand side of inequality (80) is precisely $\int_0^T \int_0^t \int_\Omega (Y_\varepsilon(s), \nabla u_\varepsilon(s))_{\mathbb{R}^N}\, dx\, ds\, dt$, which proves (76) and thus (75).

We have constructed a sequence u_ε with the following properties:

$$u_\varepsilon \in L_\infty(0, T; W_0^{1,q}(\Omega)), \qquad b(u_\varepsilon) \in L_\infty(0, T; L_2(\Omega)),$$

$$\frac{\partial b(u_\varepsilon)}{\partial t} + \varepsilon \frac{\partial u_\varepsilon}{\partial t} - \operatorname{div} D\Phi(\nabla u_\varepsilon) = f, \tag{81}$$

$$u_\varepsilon(0) = u_0, \tag{82}$$

$$b(u_\varepsilon)|_{t=0} = b(u_0). \tag{83}$$

Once again, the weak lower semicontinuity properties of the L_2, $L_{q'}$, and L_q norms imply, by virtue of (60)–(63), (68), (69), (74), and (75) that

(84) $\sqrt{\varepsilon}\, \partial u_\varepsilon / \partial t$ is bounded in $L_2(0, T; L_2(\Omega))$ independently of ε,

(85) u_ε is bounded in $L_\infty(0, T; W_0^{1,q}(\Omega))$ independently of ε,

(86) $\Phi(\nabla u_\varepsilon)$ is bounded in $L_\infty(0, T; [L_{q'}(\Omega)]^N)$ independently of ε,

(87) $b(u_\varepsilon)$ is bounded in $L_\infty(0, T; L_2(\Omega))$ independently of ε.

With the help of (84)–(87), we conclude that there exists a suitably extracted subsequence (still denoted u_ε) such that

$$\begin{aligned} &u_\varepsilon \rightharpoonup u \quad \text{weak-* in } L_\infty(0, T; W_0^{1,q}(\Omega)), \\ &\sqrt{\varepsilon}\, \frac{\partial u_\varepsilon}{\partial t} \rightharpoonup 0 \quad \text{weakly in } L_2(0, T; L_2(\Omega)), \\ &D\Phi(\nabla u_\varepsilon) \rightharpoonup Y \quad \text{weak-* in } L_\infty(0, T; [L_{q'}(\Omega)]^N), \\ &b(u_\varepsilon) \rightharpoonup \chi \quad \text{weak-* in } L_\infty(0, T; L_2(\Omega)) \end{aligned} \tag{88}$$

as ε tends to zero.

In the following section, (88) is used to pass to the limit in (81)–(83) as ε tends to zero, and the quantities χ and Y are identified.

2.4. Passing to the limit in the coercivity parameter. By virtue of (81), (84), and (86)

$$\frac{\partial b(u_\varepsilon)}{\partial t} \text{ is bounded in } L_2(0, T; W^{-1,q'}(\Omega)). \tag{89}$$

In view of (88) and (89), the limit of (81) as ε tends to zero is

$$\frac{\partial \chi}{\partial t} - \operatorname{div} Y = f, \tag{90}$$

while that of (83) is

$$\chi|_{t=0} = b(u_0). \tag{91}$$

Remark 7. The initial value of u_ε, i.e., u_0, is lost in the limiting process because of the absence of estimates on $\partial u_\varepsilon/\partial t$ that are independent of ε.

The absence of an estimate on $\partial u_\varepsilon/\partial t$ precludes the application of Aubin's lemma to u_ε. However, that lemma can be applied to $b(u_\varepsilon)$ because of (87) and (89); it implies that, as ε tends to zero,

$$b(u_\varepsilon) \to \chi \quad \text{strongly in } \mathscr{C}^0([0, T]; W^{-1,q'}(\Omega)).$$

Since u_ε converges weak-* to u in $L_\infty(0, T; W_0^{1,q}(\Omega))$ we conclude that

$$\int_0^T \int_\Omega b(u_\varepsilon(t)) u_\varepsilon(t)\, dx\, dt \to \int_0^T \int_\Omega \chi(t) u(t)\, dx\, dt \tag{92}$$

as ε tends to zero.

We introduce the functional J defined on $L_2(\Omega \times (0, T))$ as

$$J(v) = \begin{cases} \displaystyle\int_0^T \int_\Omega \Psi(v(t))\, dx\, dt & \text{if } \Psi(v) \text{ belongs to } L_1(\Omega \times (0, T)), \\ +\infty & \text{otherwise.} \end{cases}$$

In view of the properties of Ψ (cf. § 2.1), J is positive, convex, and lower semicontinuous. It is also proper, since $\Psi(0) = 0$.

A classical result of convex analysis [11, Thm. 2, p. 532] allows us to identify the subdifferential $\partial J(v)$ of J at any point v of $L_2(\Omega \times (0, T))$ as

$$\begin{aligned} \partial J(v) = \{ w \in L_2(\Omega \times (0, T)) \,|\, & w(x, t) = b(v(x, t)) \\ & \text{almost everywhere in } \Omega \times (0, T) \}. \end{aligned} \tag{93}$$

Since both u^ε and $b(u^\varepsilon)$ lie, in particular, in $L_2(\Omega \times (0, T))$, $b(u^\varepsilon)$ belongs to $\partial J(u^\varepsilon)$. Thus,

$$\int_0^T \int_\Omega b(u_\varepsilon(t)) u_\varepsilon(t)\, dx\, dt + J(w) \geqq J(u_\varepsilon) + \int_0^T \int_\Omega b(u_\varepsilon(t)) w(t)\, dx\, dt, \tag{94}$$

for any w in $L_2(\Omega \times (0, T))$. Because of (88), (92), and the weak lower semicontinuity of J, we are in a position to compute the limit of each term in inequality (94). We obtain that, for any w in $L_2(\Omega \times (0, T))$,

$$\int_0^T \int_\Omega \chi(t) u(t)\, dx\, dt + J(w) \geqq J(u) + \int_0^T \int_\Omega \chi(t) w(t)\, dx\, dt. \tag{95}$$

Inequality (95) implies that u belongs to the domain of J and that

$$\chi \in \partial J(u), \tag{96}$$

and the characterization (93) of ∂J enables us to conclude that

$$\chi = b(u). \tag{97}$$

Once again the identification of Y relies on Lemma 2. The quantities w_m, w, and Y are identified with ∇u_ε, ∇u, and Y, respectively, and (48) is satisfied with the help of the convergences (88). To show that

$$Y = D\Phi(\nabla u) \tag{98}$$

reduces to proving that

$$\begin{aligned} &\int_0^T \int_0^t \int_\Omega (Y(s), \nabla u(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\geqq \overline{\lim_{\varepsilon \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u_\varepsilon(s)), \nabla u_\varepsilon(s))_{\mathbb{R}^N} \, dx \, ds \, dt. \end{aligned} \tag{99}$$

As seen earlier, the right-hand side of (99) is the lim-sup of the right-hand side of (80) as ε tends to zero. We obtain the following:

$$\begin{aligned} &\overline{\lim_{\varepsilon \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u_\varepsilon(s)), \nabla u_\varepsilon(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\quad \leqq \int_0^T \int_0^t \int_\Omega f(s) u(s) \, dx \, ds \, dt + T \int_\Omega \Psi^*(b(u_0)) \, dx \\ &\quad - \underline{\lim_{\varepsilon \to 0}} \int_0^T \int_\Omega \Psi^*(b(u_\varepsilon(t))) \, dx \, dt. \end{aligned}$$

But Ψ^* is positive lower semicontinuous and convex on $\mathbb{R}$; thus, with the help of (97),

$$0 \leqq \int_0^T \int_\Omega \Psi^*(b(u(t))) \, dx \, dt \leqq \underline{\lim_{\varepsilon \to 0}} \int_0^T \int_\Omega \Psi^*(b(u_\varepsilon(t))) \, dx \, dt,$$

which leads to

$$\begin{aligned} &\overline{\lim_{\varepsilon \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u_\varepsilon(s)), \nabla u_\varepsilon(s)) \, dx \, ds \, dt \\ &\quad \leqq \int_0^T \int_0^t \int_\Omega f(s) u(s) \, dx \, ds \, dt + T \int_\Omega \Psi^*(b(u_0)) \, dx \\ &\quad - \int_0^T \int_\Omega \Psi^*(b(u(t))) \, dx. \end{aligned} \tag{100}$$

The right-hand side of inequality (100) is easily seen to coincide with $\int_0^T \int_0^t \int_\Omega (Y(s), \nabla u(s))_{\mathbb{R}^N} \, dx \, ds \, dt$ after multiplication of (90) by u, integration of the resulting expression over $(0, t) \times \Omega$ then over $(0, T)$, and application of Lemma 1. Inequality (99) is proved and equality (98) is established.

Recalling (88), (90), (91), (97), and (98) we conclude that there exists an element u of $L_\infty(0, T; W_0^{1,q}(\Omega))$ which satisfies (8)-(10). The proof of the existence part of Theorem 1 is now complete.

The bound on the norm of u in $L_\infty(0, T; W_0^{1,q}(\Omega))$ is a direct consequence of Remark 5 and of the weak lower semicontinuity property of the L_q norm.

2.5. Comparison result. It is now assumed that u_{01}, u_{02}, f_1, f_2 satisfy the hypotheses of Theorem 1 and that

$$b(u_{01}) \geqq b(u_{02}) \quad \text{almost everywhere on } \Omega,$$

$$f_1 \geqq f_2 \quad \text{almost everywhere on } \Omega \times (0, T).$$

Let u_1 and u_2 denote the associated solutions of (8)-(10) whose existence was established in the previous sections.

Let $\mathrm{sg}_\alpha^-(t)$ and $\mathrm{sg}_0(t)$ denote the following real-valued functions of the real variable:

$$\mathrm{sg}_\alpha^-(t) = \begin{cases} 0 & \text{if } t \geqq 0, \\ \dfrac{1}{\alpha} t & \text{if } -\alpha \leqq t \leqq 0, \\ -1 & \text{if } t \leqq -\alpha, \end{cases}$$

$$\mathrm{sg}_0(t) = \begin{cases} 0 & \text{if } t \geqq 0, \\ -1 & \text{if } t < 0. \end{cases}$$

In view of (60) and (61), the quantity $\mathrm{sg}_\alpha^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)$ is an element of $L_\infty(0, T; W_0^{1,q}(\Omega)) \cap W^{1,2}(0, T; L_2(\Omega))$ (and $b^\eta(u_\varepsilon^\eta)$ as well). Thus it is an admissible test function in (57) (with u_ε^η, f replaced by $u_{1\varepsilon}^\eta, f_1$ or $u_{2\varepsilon}^\eta, f_2$). Upon integration over the time interval $(0, t)$ of the result of the multiplication of (57) by $\mathrm{sg}_\alpha^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)$, we obtain by difference

$$\begin{aligned} &\int_0^t \int_\Omega \frac{\partial}{\partial t}[(b^\eta(u_{1\varepsilon}^\eta) + \varepsilon u_{1\varepsilon}^\eta) - (b^\eta(u_{2\varepsilon}^\eta) + \varepsilon u_{2\varepsilon}^\eta)](s)\, \mathrm{sg}_\alpha^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)(s)\, dx\, ds \\ &\quad + \int_0^t \int_\Omega (\mathrm{sg}_\alpha^-)'(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)(s)(D\Phi(\nabla u_{1\varepsilon}^\eta(s)) \\ &\qquad - D\Phi(\nabla u_{2\varepsilon}^\eta(s)), (\nabla u_{1\varepsilon}^\eta - \nabla u_{2\varepsilon}^\eta)(s))_{\mathbb{R}^N}\, dx\, ds \\ &= \int_0^t \int_\Omega (f_1 - f_2)(s)\, \mathrm{sg}_\alpha^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)(s)\, dx\, ds. \end{aligned} \tag{101}$$

Once again the derivation formula for the composition of a $W_0^{1,q}$ function by a Lipschitz function is implicitly used in (101). By virtue of the monotonicity of sg_α^- and $D\Phi$ and the positivity of $f_1 - f_2$, (101) yields

$$\int_0^t \int_\Omega \frac{\partial}{\partial t}[(b^\eta(u_{1\varepsilon}^\eta) + \varepsilon u_{1\varepsilon}^\eta) - (b^\eta(u_{2\varepsilon}^\eta) + \varepsilon u_{2\varepsilon}^\eta)](s)\, \mathrm{sg}_\alpha^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)(s)\, dx\, ds \leqq 0. \tag{102}$$

As α tends to zero, $\mathrm{sg}_\alpha^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)$ converges weak-* in $L_\infty(\Omega \times (0, T))$ to $\mathrm{sg}_0^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)$ and (102) becomes

$$\int_0^t \int_\Omega \frac{\partial}{\partial t}[(b^\eta(u_{1\varepsilon}^\eta) + \varepsilon u_{1\varepsilon}^\eta) - (b^\eta(u_{2\varepsilon}^\eta) + \varepsilon u_{2\varepsilon}^\eta)](s)\, \mathrm{sg}_0^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta)(s)\, dx\, ds \leqq 0.$$

Since $b^\eta(t) + \varepsilon t$ is monotone and takes the value zero for t equal to zero,

$$\mathrm{sg}_0^-(u_{1\varepsilon}^\eta - u_{2\varepsilon}^\eta) = \mathrm{sg}_0^-[(b^\eta(u_{1\varepsilon}^\eta) + \varepsilon u_{1\varepsilon}^\eta) - (b^\eta(u_{2\varepsilon}^\eta) + \varepsilon u_{2\varepsilon}^\eta)]$$

almost everywhere in $\Omega\times(0, T)$. The function $b^\eta(u_{i\varepsilon}^\eta)+\varepsilon u_{i\varepsilon}^\eta$ $(i=1,2)$ lies in $W^{1,2}(0, T; L_2(\Omega))$. Consequently the last inequality reads as

$$\begin{aligned}&\int_\Omega [(b^\eta(u_{1\varepsilon}^\eta)+\varepsilon u_{1\varepsilon}^\eta)-(b^\eta(u_{2\varepsilon}^\eta)+\varepsilon u_{2\varepsilon}^\eta)]^-(t)\,dx\\ &\leqq \int_\Omega [(b^\eta(u_{01})+\varepsilon u_{01})-(b^\eta(u_{02})+\varepsilon u_{02})]^-\,dx,\end{aligned}\tag{103}$$

for any t in $[0, T]$. In (103), $[\cdot]^-$ denotes the convex Lipschitz function $-\inf(\cdot, 0)$.

Since $[\ ]^-$ is convex and continuous, the convergences (63), (66), (69), and (88) together with relations (74) and (97), easily allow passing to the limit in the result of the integration of inequality (103) over the time interval $(0, T)$ as η and ε successively tend to zero. We finally conclude that

$$\int_0^T\int_\Omega [b(u_1)-b(u_2)]^-(t)\,dx\,dt \leqq T\int_\Omega [b(u_{01})-b(u_{02})]^-\,dx.\tag{104}$$

The function $[b(u_{01})-b(u_{02})]^-$ is by assumption equal to zero, which implies that for almost every (x, t) in $\Omega\times(0, T)$,

$$(b(u_1)-b(u_2))(x, t)\leqq 0.$$

The proof of Theorem 1 is now complete.

Remark 8. Note that the hypothesis on $b(u_{01})-b(u_{02})$ has not been used in the derivation of inequality (104) which thus holds true only under the assumption that f_1-f_2 is almost everywhere positive on $\Omega\times(0, T)$.

3. Proof of Theorem 2. The proof of Theorem 2 is based on the existence result given by Theorem 1. Specifically we introduce

$$u_0^n = T_n(u_0),$$

where, for any positive r, T_r is the Lipschitz function defined as

$$T_r(t)=\begin{cases} t & \text{if } |t|\leqq r,\\ r\,\mathrm{sg}\,(t) & \text{if } |t|>r.\end{cases}$$

Since T_n is a piecewise C^1 Lipschitz function and u_0 is in $W_0^{1,q}(\Omega)$, u_0^n is in $W_0^{1,q}(\Omega)$, and the derivation formula applies, from which it is easily deduced that

$$u_0^n \rightharpoonup u_0 \quad \text{weakly in } W_0^{1,q}(\Omega),\tag{105}$$

as n tends to infinity.

We also introduce a sequence f^n in $W^{1,1}(0, T; L_2(\Omega))$ such that

$$f^n \to f \quad \text{strongly in } W^{1,1}(0, T; W^{-1,q'}(\Omega))\tag{106}$$

as n tends to infinity, and we propose to study the behavior of u^n, the solution of

$$\begin{aligned}&\frac{\partial b(u^n)}{\partial t}-\operatorname{div} D\Phi(\nabla u^n)=f^n \quad \text{in } \Omega\times(0, T),\\ &u^n=0 \quad \text{on } \partial\Omega\times(0, T),\\ &b(u^n)\big|_{t=0}=b(u_0^n),\end{aligned}\tag{107}$$

as n tends to infinity. Theorem 1 ensures the existence of such a u^n in $L_\infty(0, T; W_0^{1,q}(\Omega))$ with $b(u^n)$ in $L_\infty(0, T; L_2(\Omega))\cap W^{1,\infty}(0, T; W^{-1,q'}(\Omega))$.

Furthermore the norm of u^n in $L_\infty(0, T; W_0^{1,q}(\Omega))$ is bounded by a continuous function of $\Phi(\nabla u_0^n)$ and of the norm $|||f^n|||$ of f^n in $W^{1,1}(0, T; W^{-1,q'}(\Omega))$ (cf. Theorem 1). It is then immediately deduced from (105), (106), and the properties (5) of Φ that

(108) u^n is bounded in $L_\infty(0, T; W_0^{1,q}(\Omega))$ independently of n.

Once again the boundedness of $D\Phi$ on the bounded sets of $L_q(\Omega)$ implies that

(109) $D\Phi(\nabla u^n)$ is bounded in $L_\infty(0, T; [L_{q'}(\Omega)]^N)$.

With the help of (108) and (109) we conclude that there exists a suitably extracted subsequence (still denoted u^n) such that

$$\begin{aligned} &u^n \rightharpoonup u \quad \text{weak-* in } L_\infty(0, T; W_0^{1,q}(\Omega)), \\ &D\Phi(\nabla u^n) \rightharpoonup Y \quad \text{weak-* in } L_\infty(0, T; [L_{q'}(\Omega)]^N) \end{aligned} \tag{110}$$

as n tends to infinity. Furthermore, by virtue of (106)-(109),

(111) $\dfrac{\partial b(u^n)}{\partial t}$ is bounded in $L_\infty(0, T; W^{-1,q'}(\Omega))$ independently of n.

We need to derive an estimate on $b(u^n)$ so as to be in a position to pass to the limit in (107). The function u^n is an admissible test function in the first equation of (107). Upon integration over $\Omega \times (0, t)$ of the result of the multiplication of the first equation of (107) by u^n we obtain the following estimate:

$$\int_\Omega \Psi^*(b(u^n(t)))\, dx \leqq \int_\Omega \Psi^*(b(u_0^n))\, dx + |||f^n||| \int_0^t \|u^n(s)\|_{1,q}\, ds, \tag{112}$$

where $|||f^n|||$ denotes the norm of f^n in $W^{1,1}(0, T; W^{-1,q'}(\Omega))$. Lemma 1 and the coercivity properties of Φ are implicitly used in establishing (112) (refer to (21), (27)-(29) for an identical argument).

A subsequence of $\Psi^*(b(u_0^n))$, still denoted $\Psi^*(b(u_0^n))$, converges almost everywhere to $\Psi^*(b(u_0))$. Furthermore, in view of (4) and the positivity of Ψ,

$$0 \leqq \Psi^*(b(u_0^n(x))) \leqq b(u_0^n(x))u_0^n(x) \leqq b(u_0(x))u_0(x),$$

for almost every x of Ω. Recalling Remark 2 we conclude that, as n tends to infinity

$$\int_\Omega \Psi^*(b(u_0^n))\, dx \to \int_\Omega \Psi^*(b(u_0))\, dx, \tag{113}$$

and, with the help of (112), that

(114) $\Psi^*(b(u^n))$ is bounded in $L_\infty(0, T; L_1(\Omega))$ independently of n.

In (114) we have identified $\Psi^*(b(u^n))$ with one of its subsequences.

We now make use of the following remark (cf. [1, Remark 1.2, p. 314]).

Remark 9. Let δ be an arbitrary strictly positive real number. Then

$$|b(t)| \leqq \delta \Psi^*(b(t)) + \sup_{|\sigma| \leqq 1/\delta} |b(\sigma)|$$

for every t in $\mathbb{R}$.

Remark 10. Note that Remarks 2 and 9 immediately imply that $b(u_0)$ is in fact an element of $L_1(\Omega)$.

Thus, for any strictly positive δ and almost any (x, t) in $\Omega \times (0, T)$,

(115) $$|b(u^n(x, t))| \leqq \delta \Psi^*(b(u^n(x, t))) + \sup_{|\sigma| \leqq 1/\delta} |b(\sigma)|.$$

Integration of (115) over any measurable subset Q of Ω implies, in view of (114), that, for almost any t in $(0, T)$,

$$\int_Q |b(u^n(t))|\,dx \leqq \delta\mathscr{C} + \text{mes}\,(Q) \sup_{|\sigma|\leqq 1/\delta} |b(\sigma)|, \tag{116}$$

where $\mathscr{C}$ is a generic constant independent of n.

Thus,

$$\begin{aligned} &b(u^n) \text{ is bounded in } L_\infty(0, T; L_1(\Omega)) \text{ independently of } n, \\ &b(u^n) \text{ is uniformly equi-integrable in } L_1(\Omega). \end{aligned} \tag{117}$$

We are now in a position to prove the following lemma.

LEMMA 3. *The sequence $b(u_n(t))$ lies in $\mathscr{C}^0([0, T]; L_1(\Omega))$ and as n tends to infinity*

$$b(u_n(t)) \to \chi \quad \textit{strongly in } \mathscr{C}^0([0, T]; L_1(\Omega)), \tag{118}$$

where χ is also an element of $\mathscr{C}^0([0, T]; L_1(\Omega))$.

Proof of Lemma 3. By virtue of (108) and since the embedding of $W_0^{1,q}(\Omega)$ into $L_1(\Omega)$ is compact, we conclude that there exists a measurable set Z in $(0, T)$ of zero measure such that

$$F = \{u_n(t);\ n \in \mathbb{N},\ t \in (0, T) - Z\} \text{ is relatively compact in } L_1(\Omega). \tag{119}$$

Through application of the Dunford-Pettis Theorem (see, e.g., [6, Cor. 11, p. 294]), (117) implies that

$$b(F) = \{b(u_n(t));\ n \in N,\ t \in (0, T) - Z\} \text{ is sequentially weakly relatively compact in } L_1(\Omega). \tag{120}$$

The compactness properties (119) and (120) of the sets F and $b(F)$ ensure that

$$b(F) \text{ is relatively compact in } L_1(\Omega). \tag{121}$$

Indeed, if h_n is an arbitrary sequence of $b(F)$, there exists a subsequence of h_n (still denoted by h_n) and a sequence w_n of F such that, as n tends to infinity,

$$w_n \to w \quad \text{in } L_1(\Omega) \text{ and almost everywhere in } \Omega, \tag{122}$$

$$b_n = b(w_n) \rightharpoonup h \quad \text{weakly in } L_1(\Omega). \tag{123}$$

Since b is a continuous function (122) implies the almost pointwise convergence of $b(w_n)$ to $b(w)$. The weak convergence (123) then implies the strong convergence in $L_1(\Omega)$ of the sequence $b(w_n)$ to $b(w)$ (see [6, Thm. 12, p. 295], which proves (121).

By virtue of (121), there exists a compact set K in $L_1(\Omega)$ such that

$$b(u_n(t)) - b(u_n(t')) \in K \text{ for any } n \text{ and for almost every } t \text{ and } t' \text{ in } (0, T). \tag{124}$$

In view of (111) a proper choice of s (s small enough) guarantees that

$$L^1(\Omega) \hookrightarrow W^{-1,s}(\Omega), \tag{125}$$

and

$$\frac{\partial b(u_n)}{\partial t} \text{ is bounded in } L^\infty(0, T; W^{-1,s}(\Omega)). \tag{126}$$

We now appeal to a straightforward adaptation of a classical lemma of Lions (see [8, Lemma 5.1, p. 58]), which may be proved exactly as in [12, Lemma 8, p. 84].

LEMMA 4. *Let B and Y be two Banach spaces with continuous embedding of B into Y. Let X be a compact subset of B. Then, for any strictly positive number ε, there exists a strictly positive constant C_ε such that*

$$|f|_B \leqq \varepsilon + C_\varepsilon |f|_Y \quad \textit{for any } f \textit{ in } X. \qquad \square$$

Lemma 4 is applied in our context with $B = L_1(\Omega)$, $Y = W^{-1,s}(\Omega)$, and $X = K$. Thus, with the help of (124) and (125), for any strictly positive number ε, there exists a strictly positive constant C_ε such that

$$\|b(u_n(t)) - b(u_n(t'))\|_{0,1} \leqq \varepsilon + C_\varepsilon \|b(u_n(t)) - b(u_n(t'))\|_{-1,s} \tag{127}$$

for any n and for almost every t and t' in $(0, T)$.

Estimate (126) implies that

$$\|b(u_n(t)) - b(u_n(t'))\|_{-1,s} \leqq \int_t^{t'} \left\| \frac{\partial b(u_n)(\sigma)}{\partial t} \right\|_{-1,s} d\sigma \leqq \mathscr{C}|t - t'|, \tag{128}$$

where $\mathscr{C}$ is a generic constant independent of n.

Inserting (128) into (127) leads to

$$\|b(u_n(t)) - b(u_n(t'))\|_{0,1} \leqq \varepsilon + C_\varepsilon |t - t'| \text{ for any } n \text{ and for almost every } t \text{ and } t' \text{ in } (0, T). \tag{129}$$

Estimate (129) readily yields the existence of a sequence of continuous representatives of $b(u_n(t))$ (still denoted by $b(u_n(t))$) such that for any positive number ε, there exists a strictly positive constant C_ε with

$$\|b(u_n(t)) - b(u_n(t'))\|_{0,1} \leqq \varepsilon + C_\varepsilon |t - t'| \text{ for any } n \text{ and every } t \text{ and } t' \text{ in } [0, T]. \tag{130}$$

The continuity of $b(u_n(t))$ and (121) imply that $b(u_n(t))$ is continuous on $[0, T]$ with value in a compact set of $L_1(\Omega)$ (which is independent of n). The uniform equicontinuity (130) of the sequence $b(u_n(t))$ permits the application of Ascoli's theorem, which completes the proof of Lemma 3.

It remains to prove that $\chi = b(u)$. Let ω be an arbitrary function in $L_1(\Omega \times (0, T))$. The function $T_R(b(u^n) - b(\omega))$ converges almost everywhere in $\Omega \times (0, T)$ to $T_R(\chi - b(\omega))$ and its L_∞-norm is bounded above by R. Hence

$$T_R(b(u^n) - b(\omega)) \to T_R(\chi - b(\omega)) \quad \text{strongly in } L_s(\Omega \times (0, T)), \qquad 1 \leqq s < +\infty, \tag{131}$$

as n tends to infinity. If φ denotes an arbitrary positive element of $\mathscr{C}_0^\infty(\Omega \times (0, T))$ we conclude, with the help of (110) and (131), that

$$\begin{aligned} &\lim_{n \to +\infty} \int_0^T \int_\Omega \varphi(t) T_R(b(u^n(t)) - b(\omega(t)))(u^n(t) - \omega(t))\, dx\, dt \\ &\quad = \int_0^T \int_\Omega \varphi(t) T_R(\chi(t) - b(\omega(t)))(u(t) - \omega(t))\, dx\, dt. \end{aligned} \tag{132}$$

The integrand in the left-hand side of equality (132) is always positive because b is monotone. Thus the limit is positive and we conclude that, for almost every (x, t) in $\Omega \times (0, T)$,

$$T_R(\chi(x, t) - b(\omega(x, t)))(u(x, t) - \omega(x, t)) \geqq 0. \tag{133}$$

Since R is arbitrary, (133) implies that, for almost any (x, t) in $\Omega \times (0, T)$,

$$(\chi(x, t) - b(\omega(x, t)))(u(x, t) - \omega(x, t)) \geqq 0. \tag{134}$$

A proper choice of ω in (134) then shows that

$$\chi = b(u) \quad \text{almost everywhere in } \Omega \times (0, T). \tag{135}$$

Passing to the limit in (107) is now an immediate task in view of (106), (110), (111), (118), and (135). We obtain the following:

$$\frac{\partial b(u)}{\partial t} - \operatorname{div} Y = f \quad \text{in } \Omega \times (0, T). \tag{136}$$

Since a subsequence of $b(u_0^n)$ (still denoted $b(u_0^n)$) converges almost everywhere and monotonically to $b(u_0)$, and with the help of Remark 10,

$$b(u_0^n) \to b(u_0) \quad \text{strongly in } L_1(\Omega), \tag{137}$$

as n tends to infinity. By virtue of (111), (118), and (135)

$$b(u)|_{t=0} = b(u_0). \tag{138}$$

It now remains to identify Y. The identification relies once again on Lemma 2. The quantities w_m, w, and Y are identified with ∇u^n, ∇u, and Y, respectively, and (48) is satisfied with the help of estimates (110). To show that

$$Y = D\Phi(\nabla u) \tag{139}$$

we only need to prove that

$$\begin{aligned} &\int_0^T \int_0^t \int_\Omega (Y(s), \nabla u(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\qquad \geqq \overline{\lim} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u^n(s)), \nabla u^n(s))_{\mathbb{R}^N} \, dx \, ds \, dt. \end{aligned} \tag{140}$$

As was seen earlier (cf. § 2.4) the right-hand side of (140) is the limit superior of the right-hand side of inequality (100) with f, u_0, and u, respectively, replaced by f^n, u_0^n, and u^n.

We obtain, in view of (106), (110), and (113),

$$\begin{aligned} &\overline{\lim_{\varepsilon \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u^n(s)), \nabla u^n(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\qquad \leqq \int_0^T \int_0^t \int_\Omega f(s) u(s) \, dx \, ds \, dt + T \int_\Omega \Psi^*(b(u_0)) \, dx \\ &\qquad\quad - \underline{\lim_{n \to +\infty}} \int_0^T \int_\Omega \Psi^*(b(u^n(t))) \, dx \, dt. \end{aligned}$$

But Ψ^* is positive and lower semicontinuous on $\mathbb{R}$; thus, with the help of (118), (135), and Fatou's lemma,

$$0 \leqq \int_0^T \int_\Omega \Psi^*(b(u(t))) \, dx \, dt \leqq \underline{\lim_{m \to +\infty}} \int_0^T \int_\Omega \Psi^*(b(u^n(t))) \, dx \, dt,$$

which leads to

$$\begin{aligned} &\overline{\lim_{\varepsilon \to 0}} \int_0^T \int_0^t \int_\Omega (D\Phi(\nabla u^n(s)), \nabla u^n(s))_{\mathbb{R}^N} \, dx \, ds \, dt \\ &\qquad \leqq \int_0^T \int_0^t \int_\Omega f(s) u(s) \, dx \, ds \, dt + T \int_\Omega \Psi^*(b(u_0)) \, dx \\ &\qquad\quad - \int_0^T \int_\Omega \Psi^*(b(u(t))) \, dx \, dt. \end{aligned} \tag{141}$$

The right-hand side of inequality (141) is easily seen to coincide with $\int_0^T \int_0^t \int_\Omega (Y(s), \nabla u(s))_{\mathbb{R}^N} dx\, ds\, dt$ after multiplication of (136) by u, integration of the resulting expression over $(0, t)\times\Omega$ then over $(0, T)$, and application of Lemma 1. Inequality (140) is proved and equality (139) follows.

Recalling (110), (118), and (135)-(139), we conclude that there exists an element u of $L_\infty(0, T; W_0^{1,q}(\Omega))$ which satisfies (13)-(15). The proof of the existence part of Theorem 2 is complete.

If f_1 and f_2 satisfy (12), and $f_1 - f_2$ is positive in the sense of Theorem 2, the sequences f_1^n, f_2^n introduced in (106) can be chosen such that $f_1^n - f_2^n$ is positive almost everywhere on $\Omega\times(0, T)$. According to Remark 8, inequality (104) applies to u_1^n and u_2^n. We obtain the following:

$$\int_0^T \int_\Omega [b(u_1^n) - b(u_2^n)]^-(t)\, dx\, dt \leqq T \int_\Omega [b(u_{01}^n) - b(u_{02}^n)]^-\, dx, \tag{142}$$

where u_{01}^n and u_{02}^n are the sequences associated to u_{01} and u_{02} through (105). In view of (118), (135), and (138), inequality (142) is easily seen to yield

$$\int_0^T \int_\Omega [b(u_1) - b(u_2)]^-(t)\, dx\, dt \leqq T \int_\Omega [b(u_{01}) - b(u_{02})]^-\, dx,$$

as n tends to infinity and the hypothesis on $b(u_{01}) - b(u_{02})$ permits us to conclude.

Acknowledgments. The authors thank the referee for his numerous remarks and A. Damlamian for his advice in improving the original manuscript.

REFERENCES

[1] H. W. ALT AND S. LUCKHAUS, *Quaslinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311-341.

[2] A. BAMBERGER, *Etude d'une équation doublement non linéaire*, J. Funct. Anal., 24 (1977), pp. 148-155.

[3] PH. BENILAN, *Sur un problème d'évolution non monotone dans $L_2(\Omega)$*, Internal Report, Publications mathématiques de l'Université de Besançon, France, 1975.

[4] L. BOCCARDO AND F. MURAT, *Remarques sur l'homogénéisation de certains problèmes quasi-linéaires*, Portugal. Math., 41 (1982), pp. 535-562.

[5] H. BRÉZIS AND F. BROWDER, *A property of Sobolev spaces*, Comm. Partial Differential Equations, 4 (1979), pp. 1077-1083.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Interscience, New York, 1967.

[7] O. GRANGE AND E. MIGNOT, *Sur la résolution d'une équation et d'une inéquation paraboliques non linéaires*, J. Funct. Anal., 11 (1972), pp. 77-92.

[8] J. L. LIONS, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Gauthier-Villars, Paris, 1969.

[9] M. MARCUS AND V. J. MIZEL, *Nemitsky operators on Sobolev spaces*, Arch. Rational Mech. Anal., 51 (1973), pp. 347-370.

[10] P. A. RAVIART, *Sur la résolution de certaines équations paraboliques nonlinéaires*, J. Funct. Anal., 5 (1970), pp. 299-328.

[11] R. T. ROCKAFELLAR, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525-539.

[12] J. SIMON, *Compact sets in $L_p(0, T; B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65-96.

[13] L. TARTAR, *Cours Peccot, Collège de France*, 1977 (partially written in H-Convergence, F. Murat Séminaire d'Analyse Fonctionnelle et Numérique, Alger, 1977/1978).

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

003

TRAVELING WAVE SOLUTIONS ARISING FROM A TWO-STEP COMBUSTION MODEL*

DAVID TERMAN†

Abstract. The combustion process $A \to B \to C$ may give rise to many traveling wave solutions, each traveling with a different velocity. The first reaction, $A \to B$, may produce a flame F1 with speed θ_1, while the second reaction may produce a second flame F12 with speed θ_2. If $\theta_1 < \theta_2$, then the rear flame F12 will approach the forward one, F1. One may expect that what eventually evolves is a single configuration moving with constant velocity. This corresponds to a third traveling wave solution. In this paper it is shown that if $\theta_1 < \theta_2$, then such a third wave does indeed exist.

Key words. reaction-diffusion equation, Conley index, traveling wave solution

AMS(MOS) subject classification. 35K57

1. Introduction. This paper is concerned with proving the existence of traveling wave solutions of a reaction-diffusion system which arises in the theory of combustion. The equations take the form

$$U_t = DU_{xx} + F(U) \tag{1.1}$$

where $U = (T, Y_1, \cdots, Y_{n-1}) \in \mathbb{R}^n$ and D is a positive, diagonal matrix. The traveling wave represents a combustion front in a premixed reactive gas. The components of U specify the dimensionless temperature and the mass fractions of the reactants. For a background of the physical motivation of these equations, see [2] and [13].

By a traveling wave solution of (1.1) we mean a nonconstant, bounded solution of the form $U(x, t) = U(z)$, $z = x + \theta t$. Note that a traveling wave solution satisfies the system of ordinary differential equations

$$DU'' - \theta U' + F(U) = 0.$$

We are primarily concerned with the two-step reaction process $A \to B \to C$. However, in order to motivate our results, we make a few remarks concerning the simple reaction $A \to B$. If we assume that the reaction rate is of mass action-Arrhenius form, T is the dimensionless temperature, and Y is the mass fraction of A, then (T, Y) satisfies the system

$$T_t = d_1 T_{xx} + QBY e^{-E/T}, \qquad Y_t = d_2 Y_{xx} - BY e^{-E/T}, \tag{1.2}$$

where d_1, d_2, Q, B, and E are all positive constants. We assume that (T, Y) satisfy boundary conditions of the form

$$(T(-\infty), Y(-\infty)) = (T_-, Y_-) \quad \text{and} \quad (T(\infty), Y(\infty)) = (T_+, 0). \tag{1.3}$$

A traveling wave solution then satisfies the system

$$d_1 T'' - \theta T' = -QBY e^{-E/T}, \qquad d_2 Y'' - \theta Y' = BY e^{-E/T} \tag{1.4}$$

along with the boundary conditions (1.3). To prove the existence of a traveling wave solution we must show that there exists a θ for which there exists a solution of (1.3), (1.4). Because the right-hand side (1.4) is zero only when $T = 0$ or $Y = 0$, there cannot

* Received by the editors February 25, 1986; accepted for publication (in revised form) January 6, 1987. This work was supported in part by the National Science Foundation under grant 8401719.

† Department of Mathematics, Ohio State University, Columbus, Ohio 43210.

exist any traveling wave solutions unless $T_- = 0$, which is physically not reasonable. This is often referred to as the "cold boundary difficulty" (see [13]); because the formulation (1.2) requires that the mixture react all the way in from $x = -\infty$, by the time finite x is reached the combustion would be complete. To overcome this difficulty, it is necessary to avoid reactions at a finite rate over an infinite time. A common way to do this is to introduce an ignition temperature. We replace (1.2) by

$$T_t = d_1 T_{xx} + QBYf(T), \qquad Y_t = d_2 Y_{xx} - BYf(T), \tag{1.5}$$

where $f(T)$ is continuous and satisfies the following:

(a) There exists $T_1 > 0$ such that $f(T) = 0$ for $T < T_1$;

(b) $f(T) > 0$ for $T > T_1$.

Berestycki, Nicolaenko, and Scheurer [1] prove that there does indeed exist a traveling wave solution of (1.5) which satisfies (1.3). They show that on such a solution $T(z)$ is monotone increasing, $Y(z)$ is monotone decreasing, and

$$T_+ = T_- + QY_-. \tag{1.6}$$

Note that (1.6) is easily obtained by integrating the equations in (1.4) for $-\infty < z < \infty$.

In this paper we consider the existence of traveling wave solutions arising from the reactions

$$A \xrightarrow{\text{(i)}} B \xrightarrow{\text{(ii)}} C. \tag{1.7}$$

If T is the dimensionless temperature, Y_1 the mass fraction of A, and Y_2, the mass fraction of B, then the traveling wave equations corresponding to this reaction network are

$$\begin{aligned} &d_0 T'' - \theta T' = -Q_1 Y_1 f_1(T) - Q_2 Y_2 f_2(T), \\ &d_1 Y_1'' - \theta Y_1' = Y_1 f_1(T), \\ &d_2 Y_2'' - \theta Y_2' = -Y_1 f_1(T) + Y_2 f_2(T) \end{aligned} \tag{1.8}$$

where $f_1(T) = B_1 e^{-E_1/T}$ and $f_2(T) = B_2 e^{-E_2/T}$. The constants d_0, d_1, d_2, Q_1, Q_2, B_1, B_2, E_1, and E_2 are all assumed to be positive. As before, because of the cold boundary difficulty we introduce ignition temperatures. We assume that $f_1(T)$ and $f_2(T)$ are continuous functions, and there exist positive constants T_1 and T_2 such that for $i = 1$ or 2, $f_i(T) = 0$ for $T < T_i$, and $f_i(T) > 0$ for $T > T_i$.

We assume that the unburned state

$$U_- = (T_-, Y_{1-}, Y_{2-})$$

is prescribed. Let us now imagine (1.7) as taking place in two steps, and thus producing two flames. The first reaction, $A \to B$, in (1.7) will convert the given unburned state to a partially burned state. This will produce a flame, F1, with speed say θ^1. The second reaction, $B \to C$, will then act on the product of F1 and convert the partially burned state to a completely burned state. We denote this second flame by F12 and its velocity by θ^{12}.

The above flames will proceed at different velocities. For example, if the speed of F12, which is built on the products of F1, is slower than F1 itself, then we can imagine two flames both existing, but the distance between them would be ever increasing. Now suppose that F12 is faster than F1. In this case the rear flame approaches the forward one. As it does, its effect is to heat up the forward one. We may then expect that what eventually evolves is a single configuration moving with constant velocity. This corresponds to a traveling wave solution of (1.8). In this paper we prove that if $\theta^1 < \theta^{12}$, then such a wave does indeed exist.

In order to formally state our first result we assume, for now, that $T_1 < T_2$. For convenience, we assume, without loss of generality, that the unburned state is given by

$$U_- = (T_-, Y_{1-}, Y_{2-}) = (0, 1, 1). \tag{1.9}$$

For our first result we assume that

$$0 < T_1 < Q_1 < T_2 < Q_1 + 2Q_2. \tag{1.10}$$

This condition will be needed to guarantee the existence of the simple flames, which we now define.

Suppose that $T < T_2$. Then $f_2(T) = 0$, and (T, Y_1) satisfies the reduced system

$$d_0 T'' - \theta T' = -Q_1 Y_1 f_1(T), \qquad d_1 Y_1'' - \theta Y_1' = Y_1 f_1(T). \tag{1.11}$$

This is the traveling wave system corresponding to the single reaction $A \to B$. From [1] we conclude that there exists a solution of (1.11) which satisfies

$$(T(-\infty), Y_1(-\infty)) = (0, 1) \quad \text{and} \quad (T(\infty), Y_1(\infty)) = (Q_1, 0). \tag{1.12}$$

Here we used (1.6). This wave corresponds to the flame F1. If (1.10) is satisfied, then on this solution $T(z) < Q_1 < T_2$ for all z. Hence, the solution of (1.11), (1.12) satisfies (1.8) for all z. Note that because $T(z) < T_2$, $Y_2(z)$ satisfies

$$d_2 Y_2'' - \theta Y_2' = -Y_1 f_1(T).$$

Integrating this last equation and the second equation in (1.11) for $-\infty < z < \infty$, we find that

$$-\theta(Y_2(\infty) - 1) = -\int_{-\infty}^{\infty} Y_1 f_1(T)\, dz = -\theta,$$

or $Y_2(\infty) = 2$.

We do not know that the solution of (1.11), (1.12) is unique. Let

$$\theta^1 = \sup\{\theta\text{: there exists a solution of (1.11), (1.12) with speed } \theta\}.$$

That is, θ^1 is the maximum speed of F1.

We now consider the flame F12. Along F12, $Y_1(z) = 0$. Hence, (T, Y_2) satisfies

$$d_0 T'' - \theta T' = -Q_2 Y_2 f_2(T), \qquad d_2 Y_2'' - \theta Y_2' = Y_2 f_2(T), \tag{1.13}$$

$$(T(-\infty), Y_2(-\infty)) = (Q_1, 2), \qquad (T(\infty), Y_2(\infty)) = (T_+, 0). \tag{1.14}$$

These are the traveling wave equations corresponding to the single reaction $B \to C$. From (1.6) we conclude that $T_+ = Q_1 + 2Q_2$. From [1] we conclude that there exists a solution of (1.13), (1.14). Let

$$\theta^{12} = \inf\{\theta\text{: there exists a solution of (1.13), (1.14) with speed } \theta\}.$$

In this paper we are interested in the case $\theta^1 < \theta^{12}$. We prove the following theorem.

THEOREM 1. *Assume $T_1 < T_2$, (1.10), and $\theta^1 < \theta^{12}$. Then there exists a solution of (1.8), for some speed θ, which satisfies*

$$(T, Y_1, Y_2)(-\infty) = (0, 1, 1) \quad \textit{and} \quad (T, Y_1, Y_2)(+\infty) = (Q_1 + 2Q_2, 0, 0). \tag{1.15}$$

Moreover, T, Y_1, and Y_2 are positive for all z, $T(z)$ is monotone increasing, and $Y_1(z)$ is monotone decreasing.

In this paper we only prove this theorem for the case $d_0 = d_1 = d_2 = 1$. The proof for the case of unequal diffusion constants can be found in [11]. We outline the proof in § 5 of this paper.

We now state other theorems which we can prove in a manner similar to the proof of Theorem 1. The proofs of these other results will appear in a future paper.

Assume that $T_2 < T_1$, and

$$0 < T_2 < Q_2 < T_1 < Q_1 + 2Q_2. \tag{1.16}$$

If $T < T_1$, then (T, Y_2) satisfies

$$d_0 T'' - \theta T' = -Q_2 Y_2 f_2(T), \qquad d_2 Y_2'' - \theta Y_2' = Y_2 f_2(T). \tag{1.17}$$

From [1] we conclude that there exists a solution of (1.17) which satisfies

$$(T(-\infty), Y_2(-\infty)) = (0, 1) \quad \text{and} \quad (T(\infty), Y_2(\infty)) = (Q_2, 0). \tag{1.18}$$

Let us denote this flame by F2. Because the solution of (1.17), (1.18) may not be unique we let

$$\theta^2 = \sup\{\theta: \text{there exists a solution of (1.17), (1.18) with speed } \theta\}.$$

There will be a second flame, which we denote by F21, which acts on the product of F2. We can think of this flame as converting A to B by reaction (i) in (1.7) and B thereupon being almost immediately converted to C by the faster reaction (ii). Unlike before, F21 does not correspond to the solution of a reduced system. Instead it is a solution of (1.8) together with the boundary conditions

$$(T, Y_1, Y_2)(-\infty) = (Q_2, 1, 0) \quad \text{and} \quad (T, Y_1, Y_2)(+\infty) = (Q_1 + 2Q_2, 0, 0). \tag{1.19}$$

We can then prove Theorem 2.

THEOREM 2. *Assume that* $T_2 < T_1$ *and* (1.16). *Then there exists a solution of* (1.8), (1.19) *for some* $\theta > 0$. *Moreover* T, Y_1, *and* Y_2 *are positive for all* z, T *is monotone increasing, and* $Y_1(z)$ *is monotone decreasing.*

Let

$$\theta^{21} = \inf\{\theta: \text{there exists a solution of (1.8), (1.19) with speed } \theta\}.$$

We then have Theorem 3.

THEOREM 3. *Assume that* $T_2 < T_1$, (1.16), *and* $\theta^2 < \theta^{21}$. *Then there exists a solution of* (1.8), *for some speed* θ, *which satisfies*

$$(T, Y_1, Y_2)(-\infty) = (0, 1, 1) \quad \textit{and} \quad (T, Y_1, Y_2)(+\infty) = (Q_1 + 2Q_2, 0, 0).$$

Moreover, T, Y_1, *and* Y_2 *are positive for all* z, $T(z)$ *is monotone increasing, and* $Y_1(z)$ *is monotone decreasing.*

It remains to consider the cases when (1.10) and (1.16) are not satisfied. We then have Theorem 4.

THEOREM 4. *Assume that* (1.10) *and* (1.16) *are not satisfied. Then there exists a solution of* (1.8), *for some speed* θ, *which satisfies*

$$(T, Y_1, Y_2)(-\infty) = (0, 1, 1) \quad \textit{and} \quad (T, Y_1, Y_2)(+\infty) = (Q_1 + 2Q_2, 0, 0).$$

Moreover, T, Y_1, *and* Y_2 *are positive for all* z, $T(z)$ *is monotone increasing, and* $Y_1(z)$ *is monotone decreasing.*

In the next section we outline the proof of Theorem 1. The proof is quite geometrical, and the purpose of the next section is to introduce the basic geometrical features of the proof. The proof of the theorems consists of three basic steps. The first step is to obtain a priori bounds for the solutions. These estimates are derived in § 3. The next step in the proof is to prove the theorem for the special case $d_0 = d_1 = d_2 = 1$. The proof for this case is based on the Conley index and is carried out in § 4. The last step of the proof is to continue the solution from the case $d_0 = d_1 = d_2 = 1$. As we mentioned earlier, this continuation is carried out in [11].

An essential feature in the proof of the theorem is the Conley index. Other authors have used index arguments to prove the existence of traveling wave solutions. In particular, Gardner [9] also considered a combustion problem. In that paper he introduced the method of perturbing the ignition temperature kinetics so that Conley's methods can be applied. This method is used here (see § 4B). Another important step in this paper is to attach to the traveling wave equation a flow in the θ-direction. This idea was used by Conley and Smoller [5] who considered the existence of traveling wave solutions for systems of the form $U_t = U_{xx} + \nabla F(U)$.

2. A brief outline of the proof of Theorem 1. We now briefly outline the proof of Theorem 1. We assume throughout this section that $T_1 < T_2$ and (1.10) is satisfied. The proof of Theorem 1 is quite geometrical. The purpose of this section is to introduce the basic geometric features of the proof.

The first step is to reduce (1.8) to a first order system. Let $q = T'$, $p_1 = Y_1'$, and $p_2 = Y_2'$. Then (1.8) is equivalent to the system

$$\begin{aligned} &T' = q, \\ &d_0 q' = \theta q - Q_1 Y_1 f_1(T) - Q_2 Y_2 f_2(T), \\ &Y_1' = p_1, \\ &d_1 p_1' = \theta p_1 + Y_1 f_1(T), \\ &Y_2' = p_2, \\ &d_2 p_2' = \theta p_2 - Y_1 f_1(T) + Y_2 f_2(T). \end{aligned} \tag{2.1}$$

Let $\gamma(z) = (T(z), q(z), Y_1(z), p_1(z), Y_2(z), p_2(z))$, and

$$A = (0, 0, 1, 0, 1, 0), \qquad B = (Q_1, 0, 0, 0, 2, 0), \qquad C = (Q_1 + 2Q_2, 0, 0, 0, 0, 0).$$

The flame F1 corresponds to a solution, $\gamma_1(z)$, of (2.1) which satisfies

$$\lim_{z \to -\infty} \gamma_1(z) = A \quad \text{and} \quad \lim_{z \to +\infty} \gamma_1(z) = B. \tag{2.2a}$$

The flame F12 corresponds to a solution $\gamma_{12}(z)$ of (2.1) which satisfies

$$\lim_{z \to -\infty} \gamma_{12}(z) = B \quad \text{and} \quad \lim_{z \to \infty} \gamma_{12}(z) = C. \tag{2.2b}$$

We assume, in this section, that these simple waves are unique. In particular, the wave speeds θ^1 and θ^{12} are uniquely determined. We wish to prove that if $\theta^1 < \theta^{12}$, then there exists a solution $\gamma_0(z)$ of (2.1) which satisfies

$$\lim_{z \to -\infty} \gamma_0(z) = A \quad \text{and} \quad \lim_{z \to \infty} \gamma_0(z) = 0. \tag{2.2c}$$

In each case the traveling wave solution corresponds to a trajectory in phase space which connects two critical points.

One of the key ideas in the proof of the theorems is to attach to (2.1) the equation

$$\theta' = \varepsilon(\theta - \theta_0)(\theta - \theta_1) \tag{2.3}$$

where $\theta_1 \leqq \theta_0$ and $0 \leqq \varepsilon \ll 1$. Let

$$\begin{aligned} &A_1 = (A, \theta_0), \quad B_1 = (B, \theta_0), \quad C_1 = (C, \theta_0), \\ &A_2 = (A, \theta_1), \quad B_2 = (B, \theta_1), \quad C_2 = (C, \theta_1), \\ &l_A = \{(\gamma, \theta): \gamma = A, \theta_1 \leqq \theta \leqq \theta_0\}, \\ &l_B = \{(\gamma, \theta): \gamma = B, \theta_1 \leqq \theta \leqq \theta_0\}, \\ &l_C = \{(\gamma, \theta): \gamma = C, \theta_1 \leqq \theta \leqq \theta_0\}. \end{aligned}$$

First consider the case $\varepsilon = 0$. Then l_A, l_B, and l_C correspond to lines of critical points, and $(\gamma_1(z), \theta^1)$ and $(\gamma_{12}(z), \theta^{12})$ trace out curves in phase space which "connect" these lines. This is illustrated in Fig. 1. Figure 1(a) depicts the case $\theta^1 < \theta^{12}$, and Fig. 1(b) depicts the case $\theta^1 < \theta^{12}$.

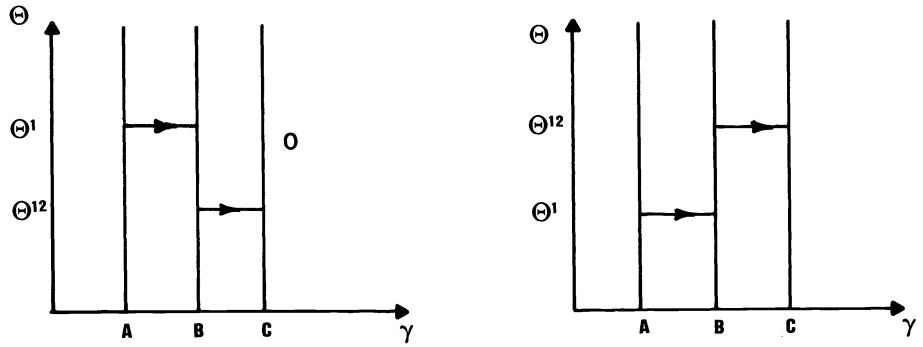
Now suppose that $\varepsilon > 0$. The lines l_A, l_B, and l_C now correspond to solutions which connect the critical points A_1 to A_2, B_1 to B_2, and C_1 to C_2, respectively. Crucial to the proof of Theorem 1 is the following result.

PROPOSITION 2.1. *For each $\varepsilon > 0$, there exists a solution $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ of (2.1), (2.3) which satisfies*

$$\lim_{z \to -\infty} (\gamma^\varepsilon(z), \theta^\varepsilon(z)) = A_1 \quad \text{and} \quad \lim_{z \to +\infty} (\gamma^\varepsilon(z), \theta^\varepsilon(z)) = C_2. \tag{2.4}$$

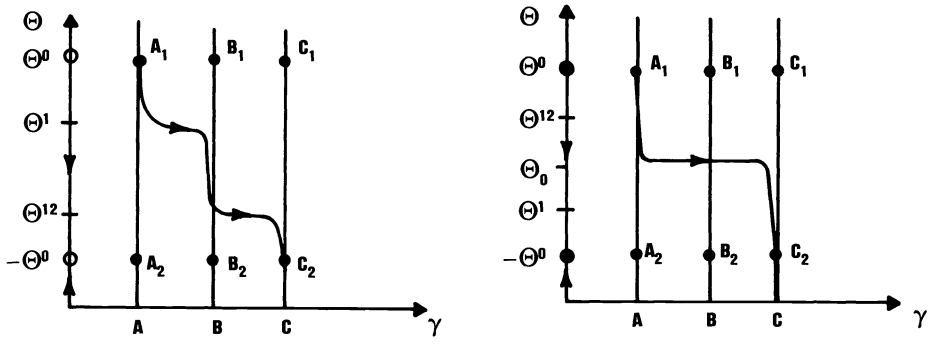
Remark. This proposition is true in both cases, $\theta^1 < \theta^{12}$ and $\theta^1 > \theta^{12}$. This result is proved later in the paper. We comment on the proof shortly. First we discuss what happens to $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ as $\varepsilon \to 0$. We shall have a priori bounds so it will be clear that at least some subsequence of $\{(\gamma^\varepsilon(z), \theta^\varepsilon(z))\}$ converges to something.

Let us, for the moment, consider the case $\theta^1 > \theta^{12}$. It is possible that for $0 < \varepsilon \ll 1$, $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ is as shown in Fig. 2(a). That is, $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ lies close to l_A for



(a) (b)

FIG. 1



(a) (b)

FIG. 2

$\theta^1 \leqq \theta \leqq \theta_0$, then lies close to $(\gamma_1(z), \theta^1)$, then lies close to l_B for $\theta^{12} \leqq \theta \leqq \theta^1$, then lies close to $(\gamma_{12}(z), \theta^{12})$, and finally lies close to l_C for $\theta_1 \leqq \theta \leqq \theta^{12}$. In the limit $\varepsilon \to 0$, the curves $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ converge to the union of the curves l_A for $\theta^1 \leqq \theta \leqq \theta_0$, $(\gamma_1(z), \theta^1)$, l_B for $\theta^{12} \leqq \theta \leqq \theta^1$, $(\gamma_{12}(z), \theta^{12})$, and l_C for $\theta_1 \leqq \theta \leqq \theta^{12}$. Hence, $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ does not converge to a traveling wave solution.

In the above paragraph we strongly used the assumption that $\theta^1 > \theta^{12}$. Since $\theta' < 0$ for $\varepsilon > 0$, it follows that if $\theta^1 < \theta^{12}$, then $(\gamma^\varepsilon(z), \theta^\varepsilon(z))$ cannot converge to a curve which is the union of $(\gamma_1(z), \theta^1)$, $(\gamma_{12}(z), \theta^{12})$ and pieces of l_A, l_B, and l_C. We prove later that the only other possibility is that there exists $\theta^0 \in (\theta_1, \theta_0)$, and a sequence $\{\varepsilon_n\}$ such that as $n \to \infty$, $\varepsilon_n \to 0$ and $\{(\gamma^{\varepsilon_n}(z), \theta^{\varepsilon_n}(z))\}$ converges to a curve which consists of three pieces. These are the following:

(a) l_A for $\theta^0 < \theta < \theta_0$;

(b) A trajectory $(\gamma_0(z), \theta^0)$ which satisfies (2.1), (2.3) for all z with $\varepsilon = 0$, $\lim_{z \to -\infty} \gamma_0(z) = A$ and $\lim_{z \to +\infty} \gamma_0(z) = C$;

(c) l_C for $\theta_1 < \theta < \theta^0$.

Then $\gamma_0(z)$ is the desired solution.

The proof of Proposition 2.1 is based on Conley's Morse Theory (see [3], [6], and [10]). First we prove the proposition for the case $d_0 = d_1 = d_2 = 1$. For general diffusion coefficients, we continue the solution. This last step is outlined in § 5.

3. A priori bounds. We now derive some a priori bounds which establish the positivity and monotonicity properties of the dependent variables. We will also find an upper bound on the speed θ of a solution. Let $\Phi(z) = (T, q, Y_1, p_1, Y_2, p_2, \theta)(z)$ be a solution of (2.1), (2.3) together with the boundary conditions

$$\lim_{z \to -\infty} \Phi(z) = A_1 \quad \text{and} \quad \lim_{z \to +\infty} \Phi(z) = C_2. \tag{3.1}$$

We assume that $\varepsilon \geqq 0$, and $\theta_0 > 0$. The motivation for considering such a solution was given in the previous section.

LEMMA 3.1. $Y_1(z) > 0$ *for all* z.

Proof. We first show that $Y_1(z) \geqq 0$ for each z. If not, there exists z_0 such that

$$Y_1(z_0) < 0 \quad \text{and} \quad Y_1'(z_0) > 0. \tag{3.2}$$

Let $\alpha = \sup\{z < z_0 \colon Y_1'(z_0) = 0\}$. Certainly α is well defined. Then $Y_1'(\alpha) = 0$, $Y_1'(z) > 0$ for $z \in (\alpha, z_0)$, and $Y_1(z) < 0$ on (α, z_0). Then

$$d_1 Y_1'' - \theta Y_1' = Y_1 f_1(T) \leqq 0 \quad \text{on } [\alpha, z_0].$$

On the other hand,

$$d_1 Y_1'' - \theta Y_1' \geqq d_1 Y_1'' - \theta_0 Y_1' \quad \text{on } [\alpha, z_0].$$

Therefore, $d_1 Y_1'' - \theta_0 Y_1' \leqq 0$ on $[\alpha, z_0]$, and $(e^{-\theta_0/d_1 Z} Y_1')' \leqq 0$ on $[\alpha, z_0]$. Integrate this last equation from α to z_0 to obtain $Y_1'(z_0) \leqq 0$. This, however, contradicts (3.2).

If $Y_1(z_0) = 0$ for some z_0, then we must have $Y_1'(z_0) = 0$. This implies that $Y_1'(z) = 0$ for all z. Since $Y_1(-\infty) = 1$, this gives a contradiction.

LEMMA 3.2. $Y_2(z) > 0$ *for all* z.

Proof. We first prove that $Y_2(z) \geqq 0$ for all z. If not, there exists z_0 such that

$$Y_2(z_0) < 0 \quad \text{and} \quad Y_2'(z_0) > 0. \tag{3.3}$$

Let $\alpha = \sup\{z < z_0;\ Y_2'(z) = 0\}$. Then $Y_2'(\alpha) = 0$, $Y_2'(z) > 0$ on (α, z_0), and $Y_2(z) < 0$ on $[\alpha, z_0]$. Therefore,

$$d_2 Y_2'' - \theta Y_2' = -Y_1 f_1(T) + Y_2 f_2(T) \leqq 0 \quad \text{on } [\alpha, z_0].$$

Therefore, $d_2 Y_2'' - \theta_0 Y_2' \leqq 0$ on $[\alpha, z_0]$, or

$$(e^{-\theta_0/d_2 z} Y_2')' \leqq 0 \quad \text{on } [\alpha, z_0].$$

Integrate this last equation from α to z_0 to obtain $Y_2'(z_0) \leqq 0$. This, however, contradicts (3.3).

If $Y_2(z_0) = 0$ for some z_0, then $Y_2'(z_0) = 0$ and $Y_2''(z_0) \geqq 0$. If $T(z_0) \leqq T_1$, then $Y_2(z) = 0$ for all z, which is impossible. If $T(z_0) > T_1$, then $d_2 Y_2'' = -Y_1 f_1(T) < 0$, which again gives a contradiction.

In a similar manner we have Lemma 3.3.

LEMMA 3.3. *$T(z) > 0$ for all z.*

Proof. The proof of this result is similar to the two just given, so we do not give the details.

LEMMA 3.4. *$Y_1'(z) \leq 0$ for all z.*

Proof. If not, then there exists z_0 such that $Y_1'(z_0) = 0$ and $Y_1''(z_0) \leqq 0$. If $T(z_0) \leqq T_1$, then $d_1 Y_1'' - \theta Y_1' = 0$. Hence, $Y_1(z) = Y_1(z_0)$ for all z, which is impossible. If $T(z_0) > T_1$, then $Y_1''(z_0) = Y_1 f_1(T) > 0$ which gives a contradiction.

LEMMA 3.5. *If $0 < T(z) < Q_1 + 2Q_2$, then $T'(z) \geqq 0$.*

Proof. If not, then because $T(+\infty) = Q_1 + 2Q_2$, there exists z_0 such that $T'(z_0) = 0$ and $T''(z_0) \geqq 0$. If $T(z_0) < T_1$, then $T(z) = T(z_0)$ for all z, which is impossible. If $T(z_0) > T_1$, then $T''(z_0) = -Q_1 Y_1 f_1(T) - Q_2 Y_2 f_2(T) < 0$, which again gives a contradiction.

Remark. The proof of this last result shows that if $T(z_0) > Q_1 + 2Q_2$ for some z_0, then there exists z_1 such that $T'(z) \geqq 0$ for $z < z_1$, and $T'(z) \leqq 0$ for $z \geqq z_1$. If this were not true then there must exist z_2 such that $T'(z_2) = 0$ and $T''(z_2) \geqq 0$. We then proceed as before.

LEMMA 3.6. *If $\varepsilon = 0$, then $T'(z) \geqq 0$ for all z.*

Proof. Consider the equation $T'' - \theta T' = -Q_1 Y_1 f_1(T) - Q_2 Y_2 f_2(T) \leqq 0$. Multiply this equation by $e^{-\theta z}$ to obtain $(e^{-\theta z} T'(z))' \leqq 0$. If we integrate this last equation from z to ∞, then we obtain the desired result.

LEMMA 3.7. *If $\varepsilon = 0$, then $0 < Y_2(z) \leqq 2$.*

Proof. If $\varepsilon = 0$, then $\theta'(z) = 0$ for all z. Suppose that $\theta(z) = \theta$. Let $u = d_1 Y_1' - \theta Y_1$, $v = d_2 Y_2' - \theta Y_2$, and $w = u + v$. Then

$$u_1'(z) = d_1 Y_1''(z) - \theta Y_1'(z) = Y_1 f_1(T) \geqq 0.$$

Hence, $u(z)$ is an increasing function. Since $u(-\infty) = \theta$ and $u(\infty) = 0$, it follows that $-\theta \leqq u(z) \leqq 0$ for all z. Moreover, $w'(z) = Y_2 f_2(T) \geqq 0$, and $w(z)$ is an increasing function. Since $w(-\infty) = -2\theta$ and $w(\infty) = 0$, it follows that $-2\theta \leqq w(z) \leqq 0$. Because $v = w - u$, it follows that $-2\theta \leqq v(z) \leqq \theta$. Hence, $d_2 Y_2' - \theta Y_2 = v \geqq -2\theta$. Multiply the left and right sides of this equation by $e^{-\theta/d_2 z}$ and integrate from z to $-\infty$ to obtain the desired result.

We conclude this section by obtaining a priori bounds on the speed θ for which there exists a solution of (2.1) which satisfies certain boundary conditions. If $\gamma(z)$ is a solution of (2.1) which satisfies either (2.2a) or (2.2c), then $\gamma(z)$ lies in $W_A^u(\theta)$, the unstable manifold of A for a particular value of θ. To obtain an upper bound on θ, we prove that if θ is sufficiently large, then each nontrivial trajectory in $W_A^u(\theta)$ becomes unbounded.

Choose L so that for $i = 1, 2$, $f_i(T) < LT$ for $0 < T < Q_1 + 3Q_2$. Let

$$\theta_0 = 2 + 2L(Q_1 + 2Q_2 + 1) + Q_1. \tag{3.4}$$

PROPOSITION 3.8. *If $\theta > \theta_0$, then each nontrivial trajectory in $W_A^u(\theta)$ is unbounded.*

Proof. Suppose that $\Phi(z) = (T, q, Y_1 p_1, Y_2, p_2)(z)$ is a solution of (2.1) with $\theta > \theta_0$ such that $\lim_{z\to-\infty} \Phi(z) = A$. Let

$$S = \{(T, q, Y_1, p_1, Y_2, p_2)\colon 0 \leqq T \leqq q \text{ or } q \leqq T \leqq 0\}.$$

First we prove that $\Phi(z) \in S$ for each z. Note that if $T < T_1$, then $f_1(T) = f_2(T) = 0$. Hence, (T, q) satisfies $T' = q$, $q' = \theta q$. This implies that $(\theta T - q)' = 0$. Because $T(-\infty) = q(-\infty) = 0$, it follows that $q = \theta T$ for all z such that $T(z) < T_1$. Since $\theta > \theta_0$, this implies that $\Phi(z) \in S$ as long as $T(z) < T_1$. In particular, there exists z_0 such that $\Phi(z) \in S$ for $z < z_0$.

We now prove that $\Phi(z) \in S$ for all $z > z_0$. If this is not true, let $z_1 = \inf\{z > z_0\colon \Phi(z) \notin S\}$. Then $\Phi(z_1) \in \partial S$. We assume that $T(z_1) > 0$. The other case is similar. Let $n = (1, -1)$. Because $\Phi(z)$ is leaving S at $z = z_1$,

$$n \cdot (T'(z_1), q'(z_1)) \geqq 0.$$

We shall prove that this is impossible. Note that $q(z_1) = T(z_1)$. Moreover, from Lemmas 3.4 and 3.7, $0 < Y_1(z) \leqq 1$ and $0 < Y_2(z) \leqq 2$. Therefore,

$$\begin{aligned} n \cdot (T'(z_1), q'(z_1)) &= q(z_1) - \theta q(z_1) + Q_1 Y_1 f_1(T) + Q_2 Y_2 f_2(T) \\ &= q(z_1)\left[1 - \theta + \frac{Q_1 Y_1 f_1(T) + Q_2 Y_2 f_2(T)}{q(z_1)}\right] \\ &= q(z_1)\left[1 - \theta + \frac{Q_1 Y_1 f_1(T) + Q_2 Y_2 f_2(T)}{T(z_1)}\right] \\ &\leqq q(z_1)[1 - \theta_0 + L(Q_1 + 2Q_2)] \\ &< 0, \end{aligned}$$

and we have the desired contradiction.

We have now shown that $\Phi(z) \in S$ for all z. Because $q(z) = T'(z)$, this implies that $T'(z) > T(z)$ for all z. Hence, $T(z) > K e^z$ for some K and the result follows.

4. The case $d_0 = d_1 = d_2 = 1$.

4A. Reduction of order. Our goal is to prove Proposition 2.1 for the case $d_0 = d_1 = d_2 = 1$, which we assume to be true throughout this section. If we multiply the second equation in (1.8) by $Q_1 + Q_2$, the third by Q_2, add the resulting equations to the first, and let

$$Z = T + (Q_1 + Q_2) Y_1 + Q_2 Y_2,$$

we find that $Z'' - \theta Z' = 0$. Since $Z'(-\infty) = Z'(\infty) = 0$, it follows that Z is constant. By assumption $Z(-\infty) = Q_1 + 2Q_2$. This implies that

$$Y_2 = \frac{1}{Q_2}[Q_1 + 2Q_2 - T - (Q_1 + Q_2) Y_1]. \tag{4A.1}$$

Plugging (4A.1) into (1.8) we obtain

$$\begin{aligned} T'' - \theta T' &= -Q_1 Y_1 f_1(T) - [Q_1 + 2Q_2 - T - (Q_1 + Q_2) Y_1] f_2(T), \\ Y_1'' - \theta Y_1' &= Y_1 f_1(T). \end{aligned} \tag{4A.2}$$

The boundary conditions (1.15) reduce to

$$(T, Y_1)(-\infty) = (0, 1), \qquad (T, Y_1)(\infty) = (Q_1 + 2Q_2, 0). \tag{4A.3}$$

4B. Perturbation of the equations. Let $P=\{(T, Y_1): T\geqq 0, Y_1\geqq 0\}$. Of course, we are only interested in values of $(T, Y_1)\in P$. The rest points of (4A.2) in P are

$$\{(T, Y_1): 0\leqq T\leqq T_1, Y_1\geqq 0\}\cup\{(T, Y_1): Y_1=0, 0\leqq T\leqq T_2\}\cup\{(Q_1+2Q_2), 0)\}.$$

This set is too large to work with. Recalling the discussion in § 2, we eventually want to reduce the problem to one in which there are only three critical points. This is done in a number of steps. We begin by perturbing (4A.2) so that the resulting system has five critical points. Let

$$F(T, Y_1)=Q_1Y_1f_1(T)+[Q_1+2Q_2-T-(Q_1+Q_2)Y_1]f_2(T).$$

Then (4A.2) can be written as

$$T''-\theta T'=F(T, Y_1), \qquad Y_1''-\theta Y_1'=Y_1f_1(T).$$

Let

$$g_1(T)=\begin{cases} T(T-T_1) & \text{if } T<Q_1,\\ 0 & \text{if } T>Q_1\end{cases}$$

and

$$g_2(T, Y_1)=-g_1(T)+Y_1\left(\frac{1}{Q_1}(Q_1-T)-Y_1\right).$$

We shall consider the perturbed system for $\varepsilon_1>0$:

(4B.1) $$T''-\theta T'=-F(T, Y_1)-\varepsilon_1 g_1(T), \qquad Y_1''-\theta Y_1'=Y_1f_1(T)-\varepsilon_1 g_2(T, Y_1).$$

Note that the rest points of (4B.1) are at the following values of (T, Y_1): $(0, 0)$, $(T_1, 0)$, $(Q_1+2Q_2, 0)$, $(0, 1)$, and $(T_1, (1/Q_1)(Q_1-T_1))$.

In what follows we consider the traveling wave system:

(4B.2)
$$\begin{aligned} T'&=q,\\ q'&=\theta q-F(T, Y_1)-\varepsilon_1 g_1(T),\\ Y_1'&=p_1,\\ p_1'&=\theta p_1+Y_1f_1(T)-\varepsilon_1 g_2(T, Y_1)\end{aligned}$$

for ε_1 small. After proving the existence of the desired connecting orbit, we let ε_1 approach zero.

4C. An isolating neighborhood. A compact set N in phase space is an isolating neighborhood if each trajectory on the boundary of N leaves N in forward or backward time. Isolating neighborhoods are important because we can assign an index (the Conley index) to the maximal invariant set inside them. Moreover, isolating neighborhoods remain isolating neighborhoods upon perturbations of the equations. In this section we construct an isolating neighborhood which contains all the trajectories of interest.

To begin with, fix δ and δ_1 positive, and let D be the set shown in Fig. 3. That is,

$$D=\left\{(T, Y_1): -\delta\leqq T\leqq -Q_1Y+2Q_2+\delta,\ Y_1\geqq\frac{1}{Q_1}(T+2\delta)\right\}$$

$$\cup\left\{(T, Y_1): -Q_1Y_1+T_1+\delta\leqq T\leqq -Q_1Y_1+Q_1+2Q_2+\delta, 0\leqq Y_1\leqq\frac{1}{Q_1}(T_1+2\delta)\right\}$$

$$\cup\{(T, Y_1): -Q_1Y_1+T_1+\delta\leqq T\leqq Q_1+2Q_2+\delta, -\delta_1\leqq Y_1\leqq 0\}.$$

Let σ_1, σ_2, σ_3, σ_4, σ_5 be the (closed) sides of D as shown in Fig. 3.

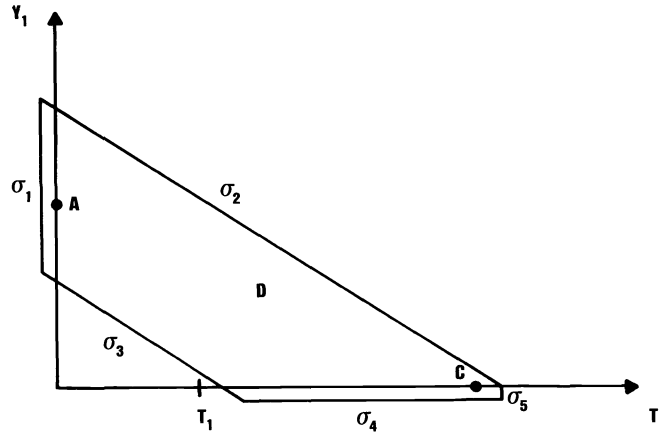


FIG. 3

LEMMA 4C.1. *There exist* δ, δ_1, *and* ε^* *such that if* $0<\varepsilon_1<\varepsilon^*$, *then* D *is positively invariant for the reaction equations*

$$\begin{aligned} T' &= F(T, Y_1)+\varepsilon_1 Q_1 g_1(T)\equiv h_1(T, Y_1),\\ Y_1' &= -Y_1 f_1(T)+\varepsilon_1 g_2(T, Y_1)\equiv h_2(T, Y_1). \end{aligned} \tag{4C.1}$$

Remark. To say that D is positively invariant means that any solution of (4C.1) which lies in D for some z_0 remains in D for all $z>z_0$.

Proof. Let

$$\begin{aligned} &\delta=\tfrac{1}{2}\min\{T_1, Q_1-T_1, 1\}\\ &M_1=\inf\{f_1(T):\ T_1+\delta<T<Q_1+2Q_2+\delta\},\\ &\hat{M}_2=T(T-T_1)\quad \text{for } T=Q_1+2Q_2+\delta,\\ &M_2=\sup\{1, \hat{M}_2\},\\ &M_3=\inf\left\{\frac{1}{2}, \frac{T_2-T_1}{2M_2^{1/2}}, Q_1\frac{M_2^{1/2}}{Q_1-T_1}, \delta Q_1 M_2^{-1/2}\right\},\\ &\varepsilon^*=\min\left\{\frac{M_1M_3}{2M_2^{1/2}}, 1\right\}\quad\text{and}\quad \delta_1=M_3M_2^{1/2}. \end{aligned}$$

Assume that $0<\varepsilon_1<\varepsilon^*$. We show that the vector field defined by the right side of (4C.1) points into D. We treat the five sides of D separately.

(a) On σ_1, $T'=\varepsilon_1 Q_1 T(T-T_1)>0$.

(b) Suppose that $(T, Y_1)\in\sigma_2$. Let $n=(1, Q_1)$ be a normal to σ_2 pointing out of D. We wish to prove that on σ_2, $n\cdot(T', Y_1')<0$ where (T', Y_1') is given by (4C.1). If $T<Q_1$, then $F_2(T)=0$. Hence, on σ_2,

$$\begin{aligned} n\cdot(T', Y_1') &= Q_1Y_1f_1(T)+\varepsilon_1Q_1T(T-T_1)-Q_1Y_1f_1(T)\\ &\quad -\varepsilon_1Q_1T(T-T_1)+\varepsilon_1Q_1Y_1\left(\frac{1}{Q_1}(Q_1-T)-Y_1\right)\\ &= -\varepsilon_1Y_1(2Q_2+\delta)<0. \end{aligned}$$

If $T\geqq Q_1$, then on σ_2,

$$\begin{aligned} n\cdot(T', Y_1') &= Q_1Y_1f_1(T)+[Q_1+2Q_2-T-(Q_1+Q_2)Y_1]f_2(T)\\ &\quad -Q_1Y_1f_1(T)+\varepsilon_1Q_1Y_1\left(\frac{1}{Q_1}(Q_1-T)-Y_1\right)\\ &= -(\delta+Q_2Y_1)f_2(T)-\varepsilon_1Y_1(\delta+2Q_2)<0. \end{aligned}$$

(c) On σ_3, $n=(-1,-Q_1)$ is an outward normal. On σ_3, $T<T_1+\delta<Q<T_2$. If $Y_1>0$, then on σ_3,

$$n\cdot(T',Y_1')=-Q_1Y_1f_1(T)-\varepsilon_1Q_1T(T-T_1)+Q_1Y_1f_1(T)-\varepsilon_1Q_1Y_1\left(\frac{1}{Q_1}(Q_1-T)-Y_1\right)$$
$$=-\varepsilon_1Q_1T(T-T_1)-\varepsilon_1Y_1(Q_1-(T+\delta))<0.$$

If $Y_1<0$, then on σ_3, $T<T_1+\delta$ and $Y_1>\delta_1$. Hence,

$$n\cdot(T',Y_1')=-\varepsilon_1Q_1T(T-T_1)-\varepsilon_1Y_1(Q_1-(T+\delta))$$
$$<-\varepsilon_1Q_1M_1+\varepsilon_1\delta_1(Q_1-T)<0.$$

(d) On σ_4, $Y_1=-\delta_1$ and $T_1+\delta<T<Q_1+2Q_2+\delta$. If $T\leqq Q_1$, then on σ_4,

$$Y_1'=-Y_1f_1(T)-\varepsilon_1T(T-T_1)+\varepsilon_1Y_1\left(\frac{1}{Q_1}(Q_1-T)-Y_1\right)$$
$$\geqq\delta_1M_1-\varepsilon_1M_2-\varepsilon_1\delta_1(1+\delta_1)$$
$$\geqq\delta_1M_1-2M_2\varepsilon^*>0.$$

If $T>Q_1$, then

$$Y_1'=-Y_1f_1(T)+\varepsilon_1Y_1\left(\frac{1}{Q_1}(Q_1-T)-Y_1\right)$$
$$\geqq\delta_1M_1-\varepsilon^*\delta_1^2>0.$$

(e) Finally, assume that $(T,Y_1)\in\sigma_5$. Then

$$T'=Q_1Y_1f_1(T)+[Q_1+2Q_2-T-(Q_1+Q_2)Y_1]f_2(T)$$
$$\leqq Q_1Y_1f_1(T)-[\delta+Q_1Y_1]f_2(T)$$
$$\leqq-[\delta-Q_1\delta_1]f_2(T)<0.$$

We now construct an isolating neighborhood for the four-dimensional flow defined by (4B.2). For $V>0$, let

$$D_1=\{(T,q,Y_1,p_1):(T,Y_1)\in D,|q|\leqq V,|p_1|\leqq V\}.$$

LEMMA 4C.2. *V can be chosen so that D_1 is an isolating neighborhood for each θ.*

Proof. Assume that $\gamma(z_0)=(T,q,Y_1,p_1)(z_0)\in\partial D_1$. We must show that $\gamma(z)$ leaves D in either backward or forward time. First assume that $|q(z_0)|<V$ and $|p_1(z_0)|<V$. Let n be an outward normal to ∂D at $(T(z_0),Y_1(z_0))$. If $\gamma(z)$ is not tangent to ∂D_1, then the result follows immediately. If $\gamma(z)$ is tangent to ∂D_1 at z_0, then $(T(z),Y_1(z))$ must be tangent to ∂D at z_0. Hence, $n\cdot(q(z_0),p_1(z_0))=0$. This implies that at $z=z_0$,

$$n\cdot(q'(z),p_1'(z))=\theta n\cdot(q(z),p(z))-n\cdot(F(T,Y_1)+\varepsilon_1g_1(T),Y_1f_1(T)-\varepsilon_1g_2(T_1,Y_1))$$
$$=-n\cdot(F(T,Y_1)+\varepsilon_1g_1(T),Y_1f_1(T)-\varepsilon_1g_2(T,Y_1))>0$$

by Lemma 4C.2. Hence, $(T(z),Y_1(z))$ is outwardly normal to ∂D at z_0. This implies the desired result.

It remains to consider the cases $|q(z_0)|=V$ and $|p(z_0)|=L$. We assume that $q(z_0)=V$. The other cases are similar. For convenience, we assume that $z_0=0$. Choose K such that

$$|F(T,Y_1)+\varepsilon_1g_1(T)|+|Y_1f_1(T)-\varepsilon_1g_2(T,Y_1)|<K\quad\text{in } D.$$

Then, if $\theta\geqq0$,

$$q'=\theta q+F(T,Y_1)+\varepsilon_1g_1(T)>-K$$

as long as $q(z)>0$. Hence, $q(z)>q(0)-Kz>V-Kz>V/2$ as long as $0<z<1$ and $V>2K$. Therefore, if $0\leqq z\leqq 1$, then $T'=q\geqq V/2$. This implies that

$$T'(1)>T(0)+\frac{V}{2}>Q_1+2Q_2+\delta$$

if V is sufficiently large. In particular, $(T(1),\ Y_1(1))\notin D$ and $\gamma(1)\notin D_1$.

Now suppose that $q(z_0)=V$ and $\theta<0$. Then

$$q'=\theta q+F(T,\ Y_1)+\varepsilon_1 g_1(T)<K$$

as long as $q(z)>0$. Hence, if $-1\leqq z\leqq 0$, then

$$q(z)>q(0)+Kz=V+Kz>\frac{V}{2}$$

as long as $-1<z<0$ and $V>2K$. Therefore, if $-1\leqq z\leqq 0$, then $T'=q>V/2$. This implies that

$$T(-1)<T(0)-\frac{V}{2}<Q_1+2Q_2+\delta-\frac{V}{2}<-\delta$$

if V is sufficiently large. In particular, $(T(1),\ Y_1(1))\notin D$ and $\gamma(1)\notin D_1$.

4D. A Morse decomposition. Let θ_0 be as in (3.4) and fix $\varepsilon>0$. Attach to (4B.2) the equation

$$\theta'=\varepsilon(\theta^2-\theta_0^2). \tag{4D.1}$$

The reason for doing this was motivated in § 2. Let

$$N=\{(\gamma,\theta):\ \gamma\in D_1,\ |\theta|\leqq\theta_0+1\}.$$

There are six critical points of (4B.2), (4D.1) in N. These are at the following values of $(T,\ q,\ Y_1,\ p_1,\ \theta)$:

$$\alpha_1=(0,0,1,0,\theta_0),\qquad \alpha_2=(0,0,1,0,-\theta_0),$$

$$\beta_1=\left(T_1,0,\frac{1}{Q_1}(T-Q_1),0,\theta_0\right),\qquad \beta_2=\left(T_1,0,\frac{1}{Q_1}(T-Q_1),0,-\theta_0\right),$$

$$\gamma_1=(Q_1+2Q_2,0,0,0,\theta_0),\qquad \gamma_2=(Q_1+2Q_2,0,0,0,-\theta_0).$$

Our immediate goal is to prove that for each $\varepsilon>0$ there exists a solution of (4B.2), (4D.1) which satisfies

$$\lim_{z\to-\infty}(\gamma(z),\theta(z))=\alpha_1\quad\text{and}\quad\lim_{z\to+\infty}(\gamma(z),\theta(z))=\gamma_2. \tag{4D.2}$$

The proof will involve the Conley index. We begin by constructing a Morse decomposition of the maximal invariant set in N.

DEFINITION (Morse decomposition). Assume that S is a compact, invariant subset of phase space. A Morse decomposition of S is a finite collection $\{M_\pi\}_{\pi\subset P}$ of subsets $M_\pi\subset S$ which are disjoint, compact, and invariant and can be ordered $\{M_1,\cdots,M_n\}$ so that for every $\gamma\in S\backslash\cup_{1\leqq j\leqq n}M_j$, there exist indices $i<j$ such that $\omega(\gamma)\in M_j$ and $\omega^*(\gamma)\in M_i$. By $\omega(\gamma)$ and $\omega^*(\gamma)$ we mean the ω-limit and ω^*-limit sets of γ, respectively.

The construction of the Morse decomposition requires a number of steps. First we define certain neighborhoods of

$$\alpha_0=(0,0,1,0)\quad\text{and}\quad\gamma_0=(Q_1+2Q_2,0,0,0).$$

Here we have given the (T, q, Y_1, p_1) coordinates of α_0 and γ_0. Let δ be as in the definition of D. Let

$$G_1 = \{(T, q): -\delta \leqq T \leqq q + \delta, 0 \leqq q \leqq V\}$$
$$\cup \{(T, q): q - \delta \leqq T \leqq \delta, -V \leqq q \leqq 0\},$$
$$G_2 = \{(Y_1, p_1): 1 - \delta \leqq Y_1 \leqq p_1 + 1 + \delta, 0 \leqq p_1 \leqq V\}$$
$$\cup \{(Y_1, p_1): p_1 + 1 - \delta \leqq Y_1 \leqq 1 + \delta, -V \leqq p_1 \leqq 0\},$$
$$G'_\alpha = \{(T, q, Y_1, p_1): (T, q) \in G_1, (Y_1, p_1) \in G_2, \text{ and } (T, Y_1) \in D\}.$$

LEMMA 4D.1. *If $\theta \geqq \theta_0 - 1$, then solutions of* (4B.2) *can only leave G'_α through the sides $|q| = V$ or $|p_1| = V$.*

Proof. We prove that on $\partial G'_\alpha \cap \{|q| < V$ and $|p_1| < V\}$ the vector field given by (4B.2) points into G'_α. If $(T, q, Y_1, p_1) \in \partial G'_\alpha$, then either $(T, q) \in \partial G_1$ or $(Y_1, p_1) \in \partial G_2$. We assume that $(T, q) \in \partial G_1 \cap \{|q| < V\}$. That proof for G_2 is similar. We treat each side of G_1 separately.

(a) Assume that $T = -\delta$ and $q > 0$. Then $T' = q > 0$.

(b) Assume that $T = \delta$ and $q < 0$. Then $T' = q < 0$.

(c) Assume that $T = q + \delta$ and $q > 0$. An outward normal to ∂G_1 is $n = (1, -1)$. If $T \leqq T_1$, then for $\theta > \theta_0 - 1$,

$$n \cdot (T', q') = q - \theta q + \varepsilon_1 T(T - T_1)$$
$$\leqq (2 - \theta_0) q < 0.$$

Here we used (3.4). If $T_1 < T \leqq Q_1$, then

$$n \cdot (T', q') = q - \theta q + Q_1 Y_1 f_1(T) + \varepsilon_1 T(T - T_1)$$
$$\leqq (2 - \theta_0) q + Q_1 Y_1 L T + \varepsilon_1 (Q_1 - T_1) T$$
$$\leqq [2 + Q_1 Y_1 L + Q_1 - Q_0] T$$
$$\leqq 0$$

where we use (3.4) and $Y_1 \leqq 1/Q_1[Q_1 + 2Q_2 + 1]$ in D.

(d) Assume that $T = q - \delta$ and $q < 0$. An outward normal to ∂G_1 at (T, q) is $n = (-1, 1)$. Moreover,

$$n \cdot (T', q') = -q + \theta q + \varepsilon_1 T(T - T_1)$$
$$< (2 - \theta_0) q < 0.$$

This completes the proof of the lemma.

In a similar fashion we construct a neighborhood of γ_0. Let

$$H_1 = \{(T, q): Q_1 + 2Q_2 - \delta \leqq T \leqq q + Q_1 + 2Q_2 + \delta, 0 \leqq q \leqq V\}$$
$$\cup \{(T, q): q + Q_1 + 2Q_2 - \delta \leqq T \leqq Q_1 + 2Q_2 + \delta, -V \leqq q \leqq 0\},$$
$$H_2 = \{(Y_1, p_1): -\delta \leqq Y_1 \leqq p_1 + \delta, 0 \leqq p_1 \leqq V\}$$
$$\cup \{(Y_1, p_1): p_1 - \delta \leqq Y_1 \leqq \delta, -V \leqq p_1 \leqq 0\},$$
$$H'_\gamma = \{(T, q, Y_1, p_1): (T, q) \in H_1 \text{ and } (Y_1, p_1) \in H_2\}.$$

As in Lemma 4D.1, we have Lemma 4D.2.

LEMMA 4D.2. *If* $\theta < -\theta_0 + 1$, *then a solution of* (4B.2) *can only enter* H'_γ *through the sides* $|q| = V$ *or* $|p_1| = V$.

We are now ready to define the Morse decomposition.

Let

$$G_\alpha = \{(T, q, Y_1, p_1, \theta): (T, q, Y_1, p_1) \in G'_\alpha, |\theta - \theta_0| \leqq 1\},$$

$$H_\gamma = \{(T, q, Y_1, p_1, \theta): (T, q, Y_1, p_1) \in G'_\alpha, |\theta + \theta_0| \leqq 1\},$$

$$N_1 = \text{cl}\,\{(\gamma, \theta): \gamma \in D_1 \backslash G'_\alpha, |\theta - \theta_0| \leqq 1\},$$

$$N_2 = G_\alpha \cup \{(\gamma, \theta): \gamma \in D_1, |\theta| \leqq \theta_0 - 1\} \cup H_\gamma,$$

$$N_3 = \text{cl}\,\{(\gamma, \theta): \gamma \in D_1 \backslash H'_\gamma, |\theta + \theta_0| \leqq 1\}.$$

By cl X we mean the topological closure of a set X. Let M_0 equal the maximal invariant set in N, and M_i, $i = 1, 2, 3$, equal the maximal invariant set in N_i. The picture we have in mind is shown in Fig. 4.

We claim that (M_1, M_2, M_3) defines a Morse decomposition of M_0.

LEMMA 4D.3. N *is an isolating neighborhood.*

Proof. Suppose that $(\gamma(x_0), \theta(z_0)) \in \partial N$. If $\theta(z_0) = \theta_0 + 1$, then $\theta'(z_0) > 0$. Therefore, $(\gamma(z), \theta(z))$ leaves N in forward time. If $\theta(z_0) = -\theta_0 - 1$, then $\theta'(z_0) > 0$. Therefore, $(\gamma(z), \theta(z))$ leaves N in backward time. If $\gamma(z) \in \partial D_1$, then because D_1 is an isolating neighborhood for each θ, it follows that $\gamma(z)$ must leave D_1 in either forward or backward time. Hence, $(\gamma(z), \theta(z))$ must leave N in forward or backward time.

LEMMA 4D.4. N_1, N_2, *and* N_3 *are all isolating neighborhoods.*

Proof. We prove the lemma for N_1; the proofs for N_2 and N_3 are similar. Suppose that $(\gamma(z_0), \theta_0(z_0)) \in \partial N_1$. If $\theta(z_0) = \theta_0 + 1$, then $\theta' > 0$ so $(\gamma(z), \theta(z))$ leaves N_1 in forward time. If $\theta = \theta_0 - 1$, then $\theta' < 0$ so, again, $(\gamma(z), \theta(z))$ leaves N_1 in forward time. If $(\gamma(z), \theta(z)) \in \partial N$, then, because N is an isolating neighborhood, $(\gamma(z), \theta(z))$ must leave N, and therefore N_1. Finally if $(\gamma(z_0), \theta(z_0)) \in \partial G_\alpha$, then by Lemma 4D.1, $(\gamma(z), \theta(z))$ must enter N_2, and therefore leave N_1, in forward time.

PROPOSITION 4D.5. (M_1, M_2, M_3) *defines a Morse decomposition for* M_0.

Proof. Trajectories that lie in N_2 cannot enter N_1 in forward time, and trajectories that lie in N_3 cannot enter N_2 in forward time. This is because on $\partial N_1 \cap \partial N_2$ the vector field points into N_2, while the vector field on $\partial N_2 \cap \partial N_3$ points into N_3.

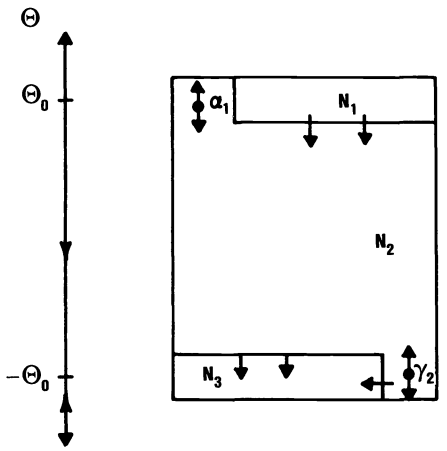


FIG. 4

Choose $\Gamma_0 \in M_0 \backslash (M_1 \cup M_2 \cup M_3)$, and let $\Gamma(z)$ be the solution of (4B.2), (4D.1) which satisfies $\Gamma(0) = \Gamma_0$. If $\Gamma_0 \in N_1$, then the above comments imply that $\Gamma(z) \notin N_2 \cup N_3$ for all $z < 0$. Hence $\omega^*(\Gamma_0) \in M_1$. Certainly $\omega(\Gamma_0) \notin M_1$, since this would imply that $\Gamma_0 \in M_1$ which we are assuming is not true. Hence, $\omega(\Gamma_0) \in M_2 \cup M_3$.

A similar argument shows that if $\Gamma_0 \in N_2$, then $\omega^*(\Gamma_0) \in M_1 \cup M_2$ and $\omega(\Gamma_0) \in M_2 \cup M_3$. If $\Gamma_0 \in N_3$, then $\omega^*(\Gamma_0) \in M_1 \cup M_2$ and $\omega(\Gamma_0) \in M_3$.

4E. Computation of the indices. We prove that $h(M_0) = h(M_1) = h(M_3) = \bar{0}$ where h is the Conley (homotopy) index and $\bar{0}$ is the Conley index of the empty set. This information will be used in the next section when we discuss $h(M_2)$ and then prove the existence of a connecting orbit.

LEMMA 4E.1. $h(M_0) = \bar{0}$.

Proof. We continue the flow (4B.2), (4D.1) to one in which the maximal invariant set in N is the empty set. Consider the following equations parametrized by λ:

$$\theta' = -\varepsilon(\theta_0^2 - \theta^2) + \lambda. \tag{4E.1λ}$$

Of course, if $\lambda = 0$ then (4E.1λ) is the same as (4D.1). The proof of Lemma 4D.3 shows that N is an isolating neighborhood for the equations (4B.2), (4E.1λ) for each $\lambda \geqq 0$. If λ_0 is sufficiently large, then $\theta' = -\varepsilon(\theta_{0\lambda}^2 - \theta^2) + \lambda_0 > 0$ for each θ. Hence (4B.2), (4E.1λ_0) does not have any bounded solutions. If I_λ is the maximal invariant set in N for the flow given by (4B.2), (4E.1λ), then $h(I_{\lambda_0}) = h(\phi) = \bar{0}$. Since $I_0 = M_0$ and I_{λ_0} are related by continuation (see [3]) it follows that $h(M_0) = \bar{0}$.

LEMMA 4E.2. $h(M_1) = \bar{0}$.

Proof. Clearly M_1 lies in the set $\{(\gamma, \theta)\colon \theta = \theta_0\}$. Let $h_1(M_1)$ equal the Conley index of M_1 considered as an isolated invariant set for the flow defined by (4B.2) with $\theta = \theta_0$. We prove that $h_1(M_1) = \bar{0}$. From elementary properties of Conley's theory, it then follows that $h(M_1) = h_1(M_1) \wedge \Sigma^1 = \bar{0}$. We shall need other properties of the Conley index.

Consider the equations (4C.1). Then $\alpha \equiv \{(T, Y_1) = (0, 1)\}$ is an attracting rest point for these equations. Choose H to be an open neighborhood of α such that $H \in D$, and on ∂H the vector field defined by (4C.1) points into H. Recall that D was defined in § 4C. It follows that $D \backslash H$ is an isolating neighborhood. Let P be the maximal invariant set in $D \backslash H$ for the equations (4C.1), and $h(P)$ equals the Conley index of P.

We wish to relate $h_1(M_1)$ with $h(P)$. To do this we need one more preliminary step. Recall the sets D_1 and G'_α. Note that $M_1 \subset N_1 \cap \{\theta = \theta_0\} = \text{cl}\,\{(\gamma, \theta)\colon \gamma \in D_1 \backslash G'_\alpha, \theta = \theta_0\}$. Let $S = \text{cl}\,(D_1 \backslash G'_\alpha)$. Lemma 4C.2 and Lemma 4D.1 imply that S is an isolating neighborhood for the flows given by (4B.2) for each $\theta \geqq \theta_0$. For $\theta \geqq \theta_0$, let $h(\theta)$ be equal to the Conley index for the maximal invariant set inside of S for the flow given by (4B.2). From the definitions, we have that $h(\theta_0) = h_1(M_1)$. From the basic Continuation Theorem [3] it follows that $h(\theta_0) = h(\theta)$ for all $\theta > \theta_0$. Hence, it remains to prove that $h(\theta) = \bar{0}$ for θ sufficiently large. From [12, Thm. I] we conclude that for θ sufficiently large, $h(\theta) = h(P)$. We now show that $h(P) = \bar{0}$.

We compute the index of P directly. Recall that P is maximal invariant set in $D \backslash H$. By Lemma 4C.1, trajectories enter $D \backslash H$ along its outer boundary. Moreover, trajectories leave $D \backslash H$ along ∂H. It is easy to imagine a vector field which enters $D \backslash H$ along its outer boundary, leaves $D \backslash H$ along ∂H, and has no invariant sets. Hence, $h(P) = h(\phi) = \bar{0}$. Of course, we could also compute $h(P)$ directly, using the isolating neighborhood $D \backslash H$.

LEMMA 4E.3. $h(M_3) = \bar{0}$.

Proof. The proof of this result is similar to the proof just given, so we do not give the details.

4F. A connecting orbit for $\varepsilon_1, \varepsilon_2 > 0$. In this section we prove the following proposition.

PROPOSITION 4F.1. *There exists a solution* $\Phi(z)$ *of* (4B.2), (4D.1) *which satisfies* $\lim_{z\to-\infty}\Phi(z)=\alpha_1$ *and* $\lim_{z\to\infty}\Phi(z)=\gamma_2$.

The first step in the proof of this result is to prove something about $h(M_2)$. Note that if S is any isolated covariant set, then $h(S)$ is the homotopy type of a certain topological space. Let $H(h(S))$ be the homology for $h(S)$ with coefficients in some fixed ring.

LEMMA 4F.2. $H(h(M_2))=0$.

Proof. Let $N_{12}=N_1\cup N_2$ and M_{12} be equal to the maximal variant set in N_{12}. Then (M_3, M_{12}) defines a Morse decomposition of M_0. It follows (see [6]) that this induces the exact sequence

$$\begin{aligned}(4F.1)\qquad &\to H^2(h(M_3))\to H^1(h(M_{12}))\to H^1(h(M_0))\to H^1(h(M_3))\\ &\to H^0(h(M_{12}))\to H^0(h(M_0))\to H^1(h(M_3))\to 0.\end{aligned}$$

Because $h(M_3)=h(M_0)=\bar{0}$, we conclude that $H(M_3)=H(M_0)=0$. From the exact sequence we conclude that $H(M_{12})=0$.

Now (M_2, M_1) defines a Morse decomposition of M_{12}. This induces another exact sequence relating $H(M_1)$, $H(M_2)$, and $H(M_{12})$. Because $H(M_1)=H(M_2)=0$, the result follows.

LEMMA 4F.3. $M_2\neq\alpha_1\cup\gamma_2$.

Proof. If this were not true, then (see [3]) we must have that $H(M_2)=H(\alpha_1)\oplus H(\gamma_2)$. Because $H(M_2)=0$, and α_1 and γ_2 are nondegenerate rest points, this is impossible.

We now return to the proof of Proposition 4F.1. Choose $\lambda_0\in M_2\backslash(\alpha_1\cup\gamma_2)$ and let $\Gamma(z)$ be the solution of (4B.2), (4D.1) which satisfies $\Gamma(0)=\lambda_0$. We prove that $\lim_{z\to-\infty}\Gamma(z)=\alpha_1$ and $\lim_{z\to\infty}\Gamma(z)=\gamma_2$. Assume that $\Gamma(z)=(\gamma(z), \theta(z))$. If $|\theta(z)|\neq\theta_0$, then $\theta'(z)<0$. It immediately follows that

$$\omega(\Gamma(z))\cup\omega^*(\Gamma(z))\subset M_2\cap\{(\gamma,\theta)\colon|\theta|=\theta_0\}.$$

To complete the proof of the proposition we show that

$$M_2\cap\{(\gamma,\theta)\colon\theta=\theta_0\}=\alpha_1\quad\text{and}\quad M_2\cap\{(\gamma,\theta)\colon\theta=-\theta_0\}=\gamma_2.$$

We only prove the first identity since the proof of the second is almost identical.

Recall the sets G_1, G_2, and G'_α defined in § 4D. Note that $M_2\subset N_2$ and

$$N_2\cap\{(\gamma,\theta)\colon\theta=\theta_0\}\subset\{(\gamma,\theta)\colon\gamma\in G'_\alpha,\ \theta=\theta_0\}\equiv G.$$

We prove that the only bounded solution in G is the rest point α_1. Suppose that $\beta_0\in G'_\alpha$, $(\beta_0,\theta_0)\neq\alpha_1$, and $(T, q, Y_1, p_1)(z)$ is the solution of (4B.2) which satisfies $(T, q, Y_1, p_1)(0)=\beta_0$. Then either $(T, q)(0)\neq(0,0)$ or $(Y_1, p_1)(0)\neq(1,0)$. Suppose that $(T, q)(0)\neq(0,0)$. The other case is similar. We prove that $(T, q)(z)$ leaves G_1 in either forward or backward time. This will imply that $(T, q, Y_1, p_1)(z)$ leaves G_1 in forward or backward time.

There are many cases to consider. For example, suppose that $T(0)>0$ and $q(0)>0$. As long as $q(z)>0$ we have that $T'(z)=q(z)>0$. We claim that as long as $(T, q)(z)\in G_1$, $q(z)>0$. If not, choose z_1 so that $q(z_1)=0$ and $q(z)>0$ for $0<z<z_1$. Then $T(z_1)>T(0)>0$, and, from the definition of G_1, if $(T, q)(z_1)\in G_1$, then $T(z_1)<T_1$.

This is due to the assumption that $\delta < \frac{1}{2}T_1$. It follows that $q'(z_1) = \theta q(z_1) - \varepsilon T(T - T_1) > 0$, which is impossible. We have now shown that as long as $(T, q)(z) \in G_1$, $T'(z) = q(z) > 0$. Because there are no rest points in G_1 for $T > 0$, we conclude that (T, q) must leave G_1.

There are other cases to consider, but their proofs are very similar to the one just given.

4G. The limit $\varepsilon_1 \to 0$. In this section we complete the proof of Proposition 2.1 if $\theta_1 = -\theta_0$ and $d_0 = d_1 = d_2 = 1$. So far we have proven that for each $\varepsilon > 0$, $0 < \varepsilon_1 < \varepsilon^*$, there exists a solution $\Phi(\varepsilon, \varepsilon_1)(z)$ of (4B.2), (4D.1) which satisfies

$$\lim_{z \to -\infty} \Phi(\varepsilon, \varepsilon_1)(z) = \alpha_1 \quad \text{and} \quad \lim_{z \to \infty} \Phi(\varepsilon, \varepsilon_1)(z) = \gamma_2.$$

Fix $\varepsilon > 0$ and let $\Phi^k(z) = \Phi(\varepsilon, 1/k\varepsilon^*)(z)$. Assume that

$$\Phi^k(z) = (T^k, q^k, Y_1^k, p_1^k, \theta^k)(z) = (\gamma^k(z), \theta^k(z)),$$

and choose the translation so that $\theta^k(0) = 0$. Since $\Phi^k(0) \in N$, a compact set, for each k it follows that some subsequence of $\{\Phi^k(0)\}$ converges to, say, Φ_0. For convenience, we assume that the entire sequence $\{\Phi^k(0)\}$ converges to Φ_0.

Let $\Phi(z)$ be the solution of (4B.2), (4D.1) with $\varepsilon_1 = 0$ which satisfies $\Phi(0) = \Phi_0$. We wish to prove that $\lim \Phi(z) = \alpha_1$ and $\lim \Phi(z) = \gamma_2$. If we then make the substitution (4A.1), this will complete the proof of Proposition 2.1 if $d_0 = d_1 = d_2 = 1$, and $\theta_1 = -\theta_0$.

Because $\varepsilon > 0$, on $\Phi(z)$, $\theta'(z) = \varepsilon(\theta^2 - \theta_0^2) < 0$. Moreover, $\Phi^k(z) \in N_2$ for all k and z. Hence,

$$\omega^*(\Phi(0)) \subset N_2 \cap \{\theta = \theta_0\} \quad \text{and} \quad \omega(\Phi(0)) \subset N_2 \cap \{\theta = -\theta_0\}.$$

We must now be more careful than in the preceding section because when $\varepsilon_1 = 0$, $N_2 \cap \{\theta = \theta_0\}$ contains uncountably many rest points. We must, therefore, understand in more detail the behavior of each $\Phi^k(z)$ near α_1 and γ_2. We use the fact that for each k, $\Phi^k(z)$ lies in both the unstable manifold at α_1 and the stable manifold at γ_2.

To prove that $\lim_{z \to -\infty} \Phi(z) = \alpha_1$, let

$$\Sigma = \{(T, q, Y_1, p_1, \theta): 0 \leqq T \leqq q \text{ or } q \leqq T \leqq 0\} \cap \{\theta \geqq \theta_0 - 1\}.$$

We prove the following lemma.

LEMMA 4G.1. *There exists z_0 such that $\Phi(z) \in \Sigma$ for $z < z_0$. This implies that $\lim_{z \to -\infty} \Phi(z) = \alpha$, because the only bounded solution in Σ is α_1.*

Proof of Lemma 4G.1. Because $\lim_{z \to -\infty} \Phi^k(z) = \alpha_1$ for each k, we may choose z_0, M so that if $k > M$ and $z < z_0$, then $\theta^k(z) > \theta_0 - 1$ and $T^k(z) < T_1$. We shall prove that if $k > M$ and $z < z_0$, then $\Phi^k(z) \in \Sigma$. This will imply the desired result.

Fix $k > M$. To complete the proof, first we show that there exists z_1 such that $\Phi_k(z) \in \Sigma$ for $z < z_1$. To prove this, note that as $z \to -\infty$, $\Phi^k(z)$ approaches α_1 tangent to the linear space spanned by the eigenvectors corresponding to the positive eigenvalues of the linearized equations at α_1. These linearized equations are the following:

$$T' = q, \qquad q' = \theta_0 q, \qquad Y_1' = p_1, \qquad p_1' = \theta_0 p_1, \qquad \theta' = 2\varepsilon\theta.$$

The positive eigenvalues are θ_0, θ_0, and 2ε, with eigenvectors $(1, \theta_0, 0, 0, 0)$, $(0, 0, 1, \theta_0, 0)$, and $(0, 0, 0, 0, 1)$, respectively. The linear subspace spanned by these eigenvectors lies in Σ as long as $\theta > \theta_0 - 1$. This proves that there exists z_1 such that $\Phi^k(z) \in \Sigma$ for $z < z_1$.

We now show that $\Phi^k(z) \in \Sigma$ for $z_1 \leqq z \leqq z_0$. If this is not true, let $z_2 = \inf\{z: \Phi^k(z) \notin \Sigma\}$. Then $\Phi^k(z)$ must be leaving Σ at z_2. We prove that this is impossible

by showing that if $\theta > \theta_0 - 1$ and $T < T_1$, then on $\partial\Sigma$ the vector field defined by (4B.2) points into Σ. We consider each side of $\partial\Sigma$ separately.

If $T(z_2) = 0$ and $q(z_2) > 0$, then $T'(z_2) = q(z_2) > 0$. If $T(z_2) = 0$ and $q(z_2) < 0$, then $T'(z_2) = q(z_2) < 0$. In both cases, the vector field points into Σ.

Suppose that $T(z_2) = q(z_2) > 0$. Let $n = (1, -1)$ be a vector outwardly normal to $\{(T, q): 0 \leqq T \leqq q\}$ at $(T(z_2), q(z_2))$. Then $n \cdot (T', q') = (1, -1) \cdot (q, \theta q) = q - \theta q < 0$ because $\theta > \theta_0 - 1 > 1$. A similar analysis holds if $T(z_2) = q(z_0) < 0$. In both cases, the vector field points into Σ. This completes the proof of the lemma.

The proof that $\lim \Phi(z) = \gamma_2$ is much easier because γ_2 is an isolated rest point. In fact, it is the only bounded solution in $N_2 \cap \{\theta = -\theta\}$.

4H. The limit $\varepsilon \to 0$. For $\varepsilon > 0$ let $\Phi_\varepsilon(z) = (T_\varepsilon, q_\varepsilon, Y_{1\varepsilon}, p_{1\varepsilon}, Y_{2\varepsilon}, p_{2\varepsilon}, \theta_\varepsilon)(z)$ be the solution of (2.1), (2.3) with $d_0 = d_1 = d_2 = 1$ which satisfies (2.4). We assume that $\theta_1 = -\theta_0$. We wish to let $\varepsilon \to 0$ and prove that if $\theta^1 < \theta^{12}$, then $\Phi_\varepsilon(z)$ converges, somehow, to the desired connecting orbit.

Choose the translation so that $T_\varepsilon(0) = \frac{1}{2}(Q_1 + 2Q_2 + T_2)$. By compactness, there exists a subsequence $\{\varepsilon_k\}$ such that $\varepsilon_k \to 0$ as $k \to \infty$, and $\Phi_{\varepsilon_k}(0)$ converges to, say, $\Phi^* = (T^*, q^*, Y_1^*, p_1^*, Y_2^*, p_2^*, \theta^*)$. Let $\Phi(z) = (T, q, Y_1, p_1, Y_2, p_2, \theta)(z)$ be the solution of (2.1), (2.3) with $\varepsilon = 0$ such that $\Phi(0) = \Phi^*$. Of course, $\theta(z) = \theta^*$ for each z.

LEMMA 4H.1. $\theta^* > 0$.

Proof. Lemma 3.5 and the remark following it imply that there exist rest points K_1 and K_2 of (2.1) such that $\lim_{z \to -\infty} \Phi(z) = (K_1, \theta^*)$ and $\lim_{z \to +\infty} \Phi(z) = (K_2, \theta^*)$. In particular, there exists τ_1, τ_2 such that $\lim_{z \to -\infty} T(z) = \tau_1$ and $\lim_{z \to +\infty} T(z) = \tau_2$. Because $T(0) = \frac{1}{2}(Q_1 + 2Q_2 + T_2)$, we conclude that

$$\tau_1 < \tfrac{1}{2}(Q_1 + 2Q_2 + T_2) < \tau_2.$$

There are no rest points of (2.1) with $T_2 < T < Q_1 + 2Q_2$ or $T > Q_1 + 2Q_2$. This implies that $\tau_1 \leqq T_2$ and $\tau_2 = Q_1 + 2Q_2$. Now $T(z)$ satisfies the following equations:

$$\text{(4H.1)} \qquad \begin{aligned} &\text{(a)} \quad T'' - \theta^* T' = -Q_1 Y_1 f_1(T) - Q_2 Y_2 f_2(T); \\ &\text{(b)} \quad T(-\infty) = \tau_1 < T_1, \text{ and } T(\infty) = Q_1 + 2Q_2. \end{aligned}$$

If $Y_1^* > 0$, then $Y_1(z) > 0$ for all z. In this case we integrate (4H.1)(a) for $-\infty < z < \infty$ to obtain

$$\theta^* > Q_1[Q_1 + 2Q_2 - \tau_1]^{-1} \int_{-\infty}^{\infty} Y_1 f_1(T)\, dz > 0.$$

If $Y_1^* = 0$, then $Y_1(z) = 0$ for all z. We again integrate (4H.1)(a) for $-\infty < z < \infty$, and use (4A.1), to obtain

$$\begin{aligned} \theta^* &= Q_2[Q_1 + 2Q_2 - \tau_1]^{-1} \int_{-\infty}^{\infty} Y_2 f_2(T)\, dz \\ &= [Q_1 + 2Q_2 - \tau_1]^{-1} \int_{-\infty}^{\infty} [Q_1 + 2Q_2 - T] f_2(T)\, dz \\ &> 0. \end{aligned}$$

Choose z_ε so that $T_\varepsilon(z_\varepsilon) = T_1$. By Lemma 3.5, $z_\varepsilon < 0$ for all $\varepsilon > 0$.

LEMMA 4H.2. *There exists δ such that if $\varepsilon < \delta$ and $z < z_\varepsilon$, then $q_\varepsilon(z) > \frac{1}{2}\theta^* T_\varepsilon(z)$, $|p_{1\varepsilon}(z)| > \frac{1}{2}\theta^* |Y_{1\varepsilon}(z) - 1|$, and $|p_{2\varepsilon}(z)| > \frac{1}{2}\theta^* |Y_{2\varepsilon} - 1|$.*

Proof. Choose δ so that if $\varepsilon<\delta$, then $|\theta_\varepsilon(0)-\theta^*|<\frac{1}{2}\theta^*$. It then follows that $\theta(z)>\frac{1}{2}\theta^*$ for $z<z_\varepsilon$. If $z<z_\varepsilon$, then $T(z)<T_1$. Hence, if $z<z_\varepsilon$, then $T'_\varepsilon=q_\varepsilon$ and $q'_\varepsilon=\theta_\varepsilon q_\varepsilon>\frac{1}{2}\theta^* q_\varepsilon$. This implies that $(\frac{1}{2}\theta^* T_\varepsilon-q_\varepsilon)'<0$. If we integrate this equation from $-\infty$ to z we find that $q_\varepsilon(z)>\frac{1}{2}\theta^* T_\varepsilon(z)$. The proofs of the other two inequalities are similar.

By compactness, there exists a subsequence $\{\varepsilon_j\}$ such that $\varepsilon_j\to 0$ as $j\to\infty$ and $\Phi_{\varepsilon j}(z_{\varepsilon j})$ converges to, say, $\Gamma^0=(T^0, q^0, Y_1^0, p_1^0, Y_2^0, p_2^0, \theta^0)$. Let $\Gamma(z)=(T, q, Y_1, p_1, Y_2, p_2, \theta)(z)$ now be the solution of (2.1), (2.3) with $\varepsilon=0$ which satisfies $\Gamma(0)=\Gamma^0$. Clearly, $\theta(z)=\theta^0$ for all z. Because $\theta'_\varepsilon(z)<0$ for each $\varepsilon>0$ and z, we conclude that $\theta^0\geqq\theta^*$. From Lemma 4H.2 we conclude the following lemma.

LEMMA 4H.3. *If* $z<0$, *then* $q(z)\geqq\frac{1}{2}\theta^* T(z)$, $|p_1(z)|\geqq\frac{1}{2}\theta^*|Y_1(z)-1|$, *and* $|p_2(z)|\geqq\frac{1}{2}\theta^*|Y_2(z)-1|$.

Corollary 4H.4 follows immediately.

COROLLARY 4H.4. $\lim_{z\to-\infty}(T, q, Y_1, p_1, Y_2, p_2)(z)=(0, 0, 1, 0, 1, 0)$.

From Lemmas 3.4 and 3.5 we have that $Y'_1(z)\leqq 0$ for all z and $T'(z)\geqq 0$ as long as $z<Q_1+2Q_2$. It follows that there must be a rest point $(T^+, q^+, Y_1^+, p_1^+, Y_2^+, p_2^+)$ such that

$$\lim_{z\to\infty}(T, q, Y_1, p_1, Y_2, p_2)(z)=(T^+, q^+, Y_1^+, p_1^+, Y_2^+, p_2^+).$$

Moreover, $T^+>T_1$. This last statement implies that $q^+=Y_1^+=p_1^+=p_2^+=0$.

LEMMA 4H.5. *Either* $(T^+, Y_2^+)=(Q_1, 2)$ *or* $(T^+, Y_2^+)=(Q_1+2Q_2, 0)$.

Proof. Suppose that $T(z)<T_2$ for all z. Then (T, Y_1) satisfies the equations

$$\text{(4H.2)}\qquad T''-\theta^0 T'+Q_1 Y_1 f_1(T)=0,\qquad Y_1''-\theta^0 Y_1'-Y_1 f_1(T)=0,$$
$$\lim_{z\to-\infty}(T, Y_1)(z)=(0, 1)\quad\text{and}\quad\lim_{z\to+\infty}(T, Y_1)(z)=(T^+, 0).$$

If we integrate the first equation in (4H.2) we find that

$$\text{(4H.3)}\qquad T^+\theta^0=Q_1\int_{-\infty}^{\infty}Y_1 f_1(T)\,dz.$$

If we integrate the second equation in (4H.2) we find that

$$\text{(4H.4)}\qquad \theta^0=\int_{-\infty}^{\infty}Y_1 f_1(T)\,dz.$$

It follows from (4H.3) and (4H.4) that $T^+=Q_1$.

Note that $Y_2''-\theta^0 Y_2'+Y_1 f_1(T)=0$. If we integrate this equation for $-\infty<z<\infty$, use (4H.4), and the fact that $Y_2(-\infty)=1$, we find that $Y_2(\infty)=2$.

Now suppose that $T(z)>T_2$ for some z. Then (T, Y_1, Y_2) satisfy the equations

$$\text{(4H.5)}\qquad \begin{gathered} T''-\theta^0 T'+Q_1 Y_1 f_1(T)+Q_2 Y_2 f_2(T)=0,\\ Y_1''-\theta^0 Y_1'-Y_1 f_1(T)=0,\\ Y_2''-\theta^0 Y_2'+Y_1 f_1(T)-Y_2 f_2(T)=0,\\ \lim_{z\to-\infty}(T, Y_1, Y_2)=(0, 1, 1),\qquad \lim_{z\to+\infty}(T, Y_1, Y_2)=(T^+, 0, 0).\end{gathered}$$

Integrate the second equation in (4H.5) for $-\infty<z<\infty$ to find that (4H.4) holds. Integrate the third equation in (4H.5) for $-\infty<z<\infty$ to obtain

$$\int_{-\infty}^{\infty}Y_2 f_2(T)\,dz=2\theta^0.$$

Finally, if we integrate the first equation in (4H.5) for $-\infty < z < \infty$ we find that $T^+ = Q_1 + 2Q_2$.

Note that if $T^+ = Q_1 + 2Q_2$, then the proof of Theorem 1 is complete for the case $d_0 = d_1 = d_2 = 1$. That is, $(T, Y_1, Y_2)(z)$ is a solution of (1.8) which satisfies (1.15). Therefore, we assume that $T^+ = Q_1$. We shall show that this implies that $\theta^1 > \theta^{12}$, thus completing the proof of Theorem 1.

Let $\{\varepsilon_j\}$ and $\{z_{\varepsilon j}\}$ be as before. That is, $\Phi_{\varepsilon J}(z_{\varepsilon J})$ converges to Γ^0 as $j \to \infty$. For each j, reparametrize the trajectories $\Phi_{\varepsilon j}(z)$ so that instead of $-\infty < z < \infty$ we have $0 < s < 1$. Call the new curves $\Lambda_j(s)$. As $j \to \infty$, the curves $\Lambda_j(s)$ will converge to a curve which we denote by

$$\Lambda(s) = (T(s), q(s), Y_1(s), p_1(s), Y_2(s), p_2(s), \theta(s)) = (\gamma(s), \theta(s)).$$

PROPOSITION 4H.6. *Suppose that* $T^+ = Q_1$. *Then there exists* $0 < s_1 < s_2 < s_3 < s_4 < 1$ *such that we have the following*:

(a) $\gamma(s) = A$ *for* $0 < s \leqq s_1$;
(b) $\theta(s) = \theta^0$ *for* $s_1 \leqq s \leqq s_2$;
(c) $\gamma(s) = B$ *for* $s_2 \leqq s \leqq s_3$;
(d) $\theta(s) = \theta^*$ *for* $s_3 \leqq s \leqq s_4$;
(e) $\gamma(s) = C$ *for* $s_4 \leqq s < 1$.

Proof. We have already shown that there exists $0 < s_1 < s_2$ such that $\gamma(s) = A$ for $0 < s \leqq s_1$, $\theta(s) = \theta^0$ for $s_1 \leqq s \leqq s_2$ and $\gamma(s_2) = B$. To determine the behavior of $\Lambda(s)$ for $s > s_2$ we must first study the local behavior of the flow (2.1), (2.3) near the points (B, θ) for $\theta^* \leqq \theta \leqq \theta_0$.

First assume that $\varepsilon = 0$ and $\theta \geqq \theta^*$. Then the linearized equation at (B, θ) has two positive eigenvalues, two negative eigenvalues, and a triple zero eigenvalue. The triple zero eigenvalue is because near (B, θ) there is a three-dimensional subset of rest points. This subset of rest points is of the form

$$\{(T, q, Y_1, p_1, Y_2, p_2, \theta): Y_1 = q_1 = p_1 = p_2 = 0\}.$$

Through (B, θ) there is a two-dimensional stable and a two-dimensional unstable manifold. We show that for j large, $\Lambda_j(s)$ approaches close to (B, θ^0) near the stable manifold at (B, θ^0) and leaves a small neighborhood of B close to the unstable manifold at (B, θ^*).

Choose new coordinates so that in the new coordinates, (2.1), (2.3) with $\varepsilon = 0$ become near (B, θ),

$$\begin{aligned} x_1' &= -\lambda_1 x_1 + g_1(\eta), & x_2' &= -\lambda_2 x_2 + g_2(\eta), \\ y_1' &= g_3(\eta), & y_2' &= g_4(\eta), \\ z_1' &= \lambda_3 z_1 + g_3(\eta), & z_2' &= \lambda_4 z_2 + g_6(\eta), \\ \theta' &= 0. \end{aligned} \tag{4H.6}$$

Here, $\eta = (x_1, x_z, y_1, y_2, z_1, z_2, \theta)$ and $g_i(\eta) = 0\|\eta\|$ for each i. Moreover, $\lambda_i > 0$ for each i, and if $\|\eta\|$ is sufficiently small and $\theta^* \leqq \theta < \theta_0$, then

$$\begin{aligned} &g_1(x_1, x_2, 0, 0, 0, 0, \theta) = g_2(x_1, x_2, 0, 0, 0, 0, \theta) = 0, \\ &g_3(0, 0, y_1, y_2, 0, 0, \theta) = g_4(0, 0, y_1, y_2, 0, 0, \theta) = 0, \\ &g_5(0, 0, 0, 0, z_1, z_2, \theta) = g_6(0, 0, 0, 0, z_1, z_2, \theta) = 0. \end{aligned} \tag{4H.7}$$

In the new coordinates $(B, \theta) = (0, 0, 0, 0, 0, 0, \theta)$. Choose $\delta_0 > 0$ so that (4H.6), (4H.7) hold for $\|\eta\| \leqq \delta_0$. For $\delta < \delta_0$, let

$$N_\delta = \{\eta: |x_1| \leqq \delta, |x_2| \leqq \delta, |z_1| \leqq \delta_0, |z_2| \leqq \delta_0, \|Y\|^2 \leqq \delta(\|z\|^2 + 1), \theta^* \leqq \theta_0\}.$$

Here, $\|Y\|^2 = y_1^2 + y_2^2$ and $\|z\|^2 = z_1^2 + z_2^2$.

We show that if δ_0 is sufficiently small, then for each $\delta < \delta_0$, trajectories can only leave N_δ through the sides $|z_1| = \delta_0$ or $|z_2| = \delta_0$. To prove this we show that on the other sides of N_δ the vector field given by (4H.6) points into N_δ.

If $x_1 = \delta$, then $x_1' = -\lambda_1 x_1 + g(\eta) < 0$ for δ_0 sufficiently small. If $x_1 = -\delta$, then $x_1' = -\lambda_1 x_1 + g_1(\eta) > 0$ for δ_0 sufficiently small. A similar analysis shows that if $|x_2| = \delta$, then the vector field given by (4H.6) points into N_δ.

Now supppose that $\|Y\|^2 = \delta(\|z\|^2 + 1)$. Let

$$n = (0, 0, 2y_1, 2y_2, -2\delta^2 z_1, -2\delta^2 z_2, 0)$$

be a vector outwardly normal to N_δ at η. If x is the vector field given by the right side of (4H.6), then

$$n \cdot x = 2y_1 g_3(\eta) + 2y_2 g_4(\eta) - 2\delta^2 \lambda_3 z_1^2 - 2\delta^2 \lambda_4 z_2^2 - 2\delta^2 z_1 g_5(\eta) - 2\delta^2 z_2 g_6(\eta) < 0$$

if δ_0 is sufficiently small. This is what we wished to prove.

For $\delta < \delta_0$, let $E_\delta = \{\eta: |z_1| = \delta_0 \text{ or } |z_2| = \delta_0\} = N_\delta$. We have shown that any solution of (2.1), (2.3) with $\varepsilon = 0$ can only leave N_δ through E_δ. By continuity of the solutions of an ordinary differential equation with respect to a parameter we conclude that for each $\delta < \delta_0$ there exists ε_δ such that if $0 < \varepsilon < \varepsilon_\delta$ and $\Phi(z)$ is a solution of (2.1), (2.3) which lies in N_δ for some z, then $\Phi(z)$ can only leave N_δ through E_δ.

We have shown that $\lim_{z \to \infty} \Gamma(z) = (B, \theta^0)$. Hence, for each $\delta, \Gamma(z) \in N_\delta$ for z sufficiently large. Another way to say this is that for each δ, $\Lambda(s) \in N_\delta$ for s sufficiently close to s_2. This implies that given δ, there exists K_δ such that if $k > K_\delta$, then $\Lambda_k(s) \in N_\delta$ for some s. Now $\Phi_\varepsilon(z) \to C_2$ as $z \to \infty$ for each $\varepsilon > 0$. Therefore, if $k > K_\delta$, then $\Lambda_k(s)$ must leave N_δ. Choosing K_δ larger, if necessary, we conclude that if $k > K_\delta$, then $\Lambda_k(s)$ leaves N_δ through E_δ. Let $\delta_k = 1/k\delta_0$. For $k > K_\delta$, choose s_k so that $\Lambda_k(s_k) \in E_{\delta_k}$. Then $\{\Lambda_k(s_k)\}$ will converge to a point $\Lambda_0 \in \bigcap_k E_{\delta_k}$. However, $\bigcap E\delta_k$ lies in $W^u_{(B,\theta)}$, the unstable manifold of (2.1), (2.3) with $\varepsilon = 0$ at the point (B, θ) for some $\theta \in (\theta^*, \theta_0)$. Let $s_3 = \sup\{s > s_2: \Lambda(s) = (B, \theta) \text{ for some } \theta\}$, and choose $\sigma_0 > s_3$ so that $\Lambda(\sigma_0) = \Lambda_0$. Note that $\sigma_0 = \lim_{k \to \infty} s_k$. Let $\theta_* = \theta(s_3)$. We have now shown that $\gamma(s) = B$ for $s_2 \leqq s \leqq s_3$ and $\Lambda(s) \in W^u_{(B,\theta_*)}$ for $s \in (s_3, \sigma_0)$.

We must now analyze what happens for $s > \sigma_0$. Let $\hat{\Gamma}(z)$ be the solution of (2.1), (2.3) with $\varepsilon = 0$ such that $\hat{\Gamma}(0) = \Lambda_0$. From Lemma 3.5 and the remark following it, there exists a critical point K_0 of (2.1), (2.3) with $\varepsilon = 0$ and $\theta = \theta_*$ such that $\lim_{z \to \infty} \hat{\Gamma}(z) = K_0$. We shall prove that $K_0 = (C, \theta_*)$. This will complete the proof of Proposition 4H.6. From the definitions it is clear that $\theta_* = \theta^*$.

We assume that $\hat{\Gamma}(z) = (T, q, Y_1, p_1, Y_2, p_2, \theta)(z)$. Because the Y_1 component of B is zero, and $Y_1(z)$ is nonincreasing (Lemma 3.4), it follows that $Y_1(z) = 0$ for all z. Hence, $(T(z), Y_2(z))$ satisfies the equations

$$\text{(4H.8)} \qquad T'' - \theta_* T' + Q_2 Y_2 f_2(T) = 0, \qquad Y_2'' - \theta_* Y_2' - Y_2 f_2(T) = 0,$$

$\lim_{z \to +\infty} (T(z), Y_2(z)) = (Q_1, 2)$, and $\lim_{z \to \infty} (T(z), Y_2(z)) = (r_1, r_2)$ for some r_1, r_2. Of course, we wish to prove that $(r_1, r_2) = (Q_1 + 2Q_2, 0)$.

For $Q_1 \leqq T < T_2$, $f_2(T) = 0$. Hence, while $Q_1 \leqq T(z) \leqq T_2$, $T(z)$ satisfies $T'' - \theta_* T' = 0$. Integrate this equation from $-\infty$ to z to find that $T' = \theta_*(T - Q_1) > 0$. Hence, $T(z)$ is strictly increasing while $Q_1 \leqq T(z) \leqq T_2$. This implies that $r_1 > T_2$. This, however, implies that $r_2 = 0$.

Integrate the second equation in (4H.8) for $-\infty < z < \infty$ to obtain

$$\theta_* = \frac{1}{2}\int_{-\infty}^{\infty} Y_2 f_2(T)\, dz.$$

Integrate the first equation in (4H.8) for $-\infty < z < \infty$ to obtain

$$-\theta_*(r_1 - Q_1) = -Q_2 \int_{-\infty}^{\infty} Y_2 f_2(T)\, dz = -2\theta_* Q_2,$$

or $r_1 = Q_1 + 2Q_2$. The proof of Proposition 4H.6 is now complete.

We are now ready to complete the proof of Theorem 1. From Lemma 4H.5, either $(T^+, Y_2^+) = (Q_1, 2)$ or $(T^+, Y_2^+) = (Q_1 + 2Q_2, 0)$. We have already seen that if $(T^+, Y_2^+) = (Q_1 + 2Q_2, 0)$, then the proof of Theorem 1 is complete. So assume that $(T^+, Y_2^+) = (Q_1, 2)$, and consider Proposition 4H.6. Note that $\gamma(s)$ for $s_1 \leqq s \leqq s_2$ corresponds to a solution of (1.11), (1.12) with $\theta = \theta^0$. From the definition of θ^1, we conclude that $\theta^1 > \theta^0$. On the other hand, $\gamma(s)$ for $s_3 \leqq s \leqq s_4$ corresponds to a solution of (1.13), (1.14) with speed θ^*. From the definition of θ^{12} we conclude that $\theta^* > \theta^{12}$. Because $\theta^0 > \theta^*$ we have that $\theta^1 > \theta^{12}$, and the proof of Theorem 1 is complete.

5. Distinct diffusion constants. We briefly describe how to prove Theorem 1 for the case of distinct diffusion constants. The major difference in the proof is the derivation of the a priori bounds. For the case of equal diffusion constants we constructed an isolating neighborhood N. This gave us the desired bounds because the solution had to lie in N. In [11] we derive the a priori bounds for the case of distinct diffusion constants.

Once we have the a priori bounds, the basic outline of the proof of Theorem 1 is the same. The key idea is the geometric construction described in § 2. Proposition 2.1 is proved by defining a homotopy between the equations with equal diffusion constants and those with distinct diffusion constants. There is some difficulty here because for the case of equal diffusion constants, we assume that $\theta_1 = -\theta_0$ in (2.3), while the a priori bounds for distinct diffusion constants only hold for $\theta > 0$. For this reason, the homotopy is carried out in two steps. First we assume that the diffusion constants are equal, and continue the solution of (2.1), (2.3), (2.4) with $\theta_1 < 0$ to a solution with $\theta_1 > 0$. We then fix θ_1 and then homotopy the diffusion constants. These continuations are quite technical, and are carried out in [11, §§ 5I and 6].

Acknowledgment. I thank R. Gardner for carefully reading the original manuscript of this paper. His many suggestions greatly helped to simplify and clarify the proofs.

REFERENCES

[1] H. BERESTYCKI, B. NICOLAENKO, AND B. SCHEURER, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16 (1985), pp. 1207-1242.

[2] J. BUCKMASTER AND G. S. S. LUDFORD, *Theory of Laminar Flames*, Cambridge University Press, London, 1982.

[3] C. CONLEY, *Isolated Invariant Sets and The Morse Index*, CBMS Regional Conference Series in Mathematics, 38, American Mathematical Society, Providence, RI, 1978.

[4] C. CONLEY AND R. GARDNER, *An application of the generalized Morse index to traveling wave solutions of a competitive reaction-diffusion model*, Indiana Univ. Math. J. 33 (1984), pp. 319-343.

[5] C. CONLEY AND J. SMOLLER, *Algebraic and topological invariants for reaction-diffusion equations*, in Systems of Partial Differential Equations, John Ball, ed., NATO ASI Series 11, 1983, pp. 3-24.

[6] C. CONLEY AND E. ZEHNDER, *Morse-type index theory for flows and periodic solutions for Hamiltonian equations*, Comm. Pure Appl. Math., 37 (1984), pp. 207-253.

[7] P. C. FIFE AND B. NICOLAENKO, *The singular perturbation approach to flame theory with chain and competing reactions*, in Ordinary and Partial Differential Equations, W. N. Everitt and B. D. Sleemans, eds., Lecture Notes in Mathematics 9645, Springer-Verlag, Berlin, New York, 1982, pp. 232-250.

[8] ———, *Asymptotic flame theory with complex chemistry*, in Nonlinear Partial Differential Equations, J. Smoller, ed., Contemporary Mathematics 17, American Mathematical Society, Providence, RI, 1983, pp. 235-255.

[9] R. GARDNER, *On the detonation of a combustible gas*, Trans. Amer. Math. Soc., 277 (1983), pp. 431-468.

[10] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, New York, 1983.

[11] D. TERMAN, *Traveling wave solutions arising from a combustion model*, University of Minnesota I.M.A. Preprint Series 216, 1986.

[12] ———, *Connection problems arising from nonlinear diffusion equations*, in Proc. Microconference on Nonlinear Diffusion, J. Serrin, ed., to appear.

[13] F. WILLIAMS, *Combustion Theory*, Addison-Wesley, Reading, MA, 1963.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

004

SOME EXPLICIT FORMULAE FOR THE SINGULAR VALUES OF CERTAIN HANKEL OPERATORS WITH FACTORIZABLE SYMBOL*

CIPRIAN FOIAS†, ALLEN TANNENBAUM‡, AND GEORGE ZAMES§

Abstract. In this paper a determinantal formula is written that allows one to compute the singular values of Hankel operators, the L^∞-symbols of which are of the form $\bar{m}w$ for $w \in H^\infty$ rational and $m \in H^\infty$ inner. (All of the Hardy spaces are defined on the unit circle in the usual way.) This is related, moreover, to some problems from control and systems theory.

Key words. Hankel operator, compressed shift, discrete spectrum, singular values, H^∞-optimization

AMS(MOS) subject classifications. 47A20, 93B35

1. Introduction. In the past few years there has been a substantial literature devoted to the computation of the norm and, more generally, singular values of Hankel operators, the L^∞-symbol of which is of the form $\bar{m}w$ for $w \in H^\infty$ rational and $m \in H^\infty$ inner. (All of our Hardy spaces will be defined on the open unit disc D following the standard conventions of [9].) A partial reference list of this work can be found in the monograph [5].

A strong motivation for studying this problem comes from control engineering, e.g., from H^∞-optimal sensitivity theory, and from Hankel norm approximation problems in system design. (Once again we refer the interested reader to [5] for the relevant physical background.)

This paper is based on the authors' previous work [2]–[4] and [10]. We put these ideas together here, and write an elementary procedure for the computation of the singular values of the above operators based on a determinantal formula, which we derive in § 3 (see (3.8)).

More precisely, let us take our point of view from [8] and [9]. Given $m \in H^\infty$ inner and nonconstant let $H^2 \ominus mH^2$ denote the orthogonal complement of mH^2 in H^2, and let $P: H^2 \to H^2 \ominus mH^2 =: H$ denote the orthogonal projection. Given $w \in H^\infty$, $M_w: H^2 \to H^2$ denotes the operator induced by multiplication by w. We now set $w(T) := PM_w | H$. In particular, $T := PS | H$ for $S: H^2 \to H^2$, the unilateral right shift. (T is called the *compressed shift*.) Then it is completely standard to show [7] that in order to solve the aforementioned Hankel singular value problem, we can equivalently find the singular values of $w(T)$.

In point of fact, the more general problem we solve in this paper is the rather explicit computation of the discrete spectrum of operators of the form $w(T)w(T)^*$, where $T \in C_0(1)$ and $w \in H^\infty$ is rational. Recall from [9] that a contraction T on a Hilbert space H is of class $C_0(1)$ if $T^n \to 0$ and $T^{*n} \to 0$ strongly, and the operators (the squares of the "defect" operators) $I - TT^*$ and $I - T^*T$ have rank 1. Such operators appear in great abundance in mathematics and in a number of physical problems. See the recent treatise [6].

* Received by the editors April 20, 1987; accepted for publication January 22, 1988. This research was supported in part by grants from the Research Fund of Indiana University, National Science Foundation grant ECS-8704047, Air Force Office of Scientific Research grant AFOSR-88-0020, and the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

‡ Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455, and Department of Mathematics, Ben-Gurion University, Beer Sheva, Israel.

§ Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada H3A 2K6.

Although the methods we employ in this note are basically operator theoretic, there is an important algebraic constituent as well, which allows us via a determinantal formula (see (3.8) below) to explicitly determine certain invertible elements of the noncommutative ring of operators $\mathbf{C}[T, T^*]$. We believe this is the main mathematical contribution of this paper.

Hopefully, some other uses will be found for our methods, both from the theoretical and applied points of view. In particular, we believe it would be very interesting to digitally implement some of the formulae given in § 3.

2. Problem statement and preliminary results. As we noted in the Introduction, we are interested in determining the discrete spectrum of $w(T)w(T)^*$ for $w \in H^\infty$ rational, and T the compressed shift associated to the nonconstant inner function $m \in H^\infty$. Recall that the discrete spectrum of a bounded self-adjoint operator B, denoted by $\sigma_d(B)$, consists of the isolated points of the spectrum $\sigma(B)$, which are eigenvalues of finite multiplicity. For such an operator B, the essential spectrum (denoted by $\sigma_e(B)$) is the complement of $\sigma_d(B)$ in the spectrum $\sigma(B)$. (See, e.g., [6] for a more detailed discussion.) We should also mention that for the compressed shift T, we have that $\sigma_d(T) = \sigma(T) \cap D$ (the eigenvalues of finite multiplicity), while for the essential spectrum we have $\sigma_e(T) = \sigma(T) \cap \partial D$, where D denotes the open unit disc and ∂D the unit circle (see [6], [9]).

Now in order to avoid some (minor) technical difficulties we will assume throughout this paper that w is *not* a constant multiple of a Blaschke product. Indeed, in the event w is a constant times a (finite) Blaschke product, all of the "s-numbers" of the Hankel associated to $\bar{m}w$ will be equal to $\|w\|_\infty$ when $\deg m > \deg w$ (see [1], [6]). Thus the interesting case of irrational m is easily solved. Moreover, the case of $\deg m \leqq \deg w$ can be handled using classical Nevanlinna–Pick interpolation theory.

We now express $w = p/q$ as a ratio of relatively prime polynomials, and we set $n := \max\{\deg p, \deg q\}$. For $\rho \in \mathbf{R}$, let

$$P_\rho := q(T)\left(I - \frac{1}{\rho^2} w(T)w(T)^*\right)q(T)^* = q(T)q(T)^* - \frac{1}{\rho^2} p(T)p(T)^*.$$

We can clearly write

$$P_\rho = \sum_{k,j=0}^{n} C_{kj}^\rho T^k T^{*j}$$

for some constants C_{kj}^ρ with the property

$$C_{kj}^\rho = \bar{C}_{jk}^\rho. \tag{1}$$

For $z \in \mathbf{C}$, define

$$\phi_\rho(z, \bar{z}) := \sum_{k,j=0}^{n} C_{kj}^\rho z^k \bar{z}^j \tag{2}$$

and note that

$$\phi_\rho(T, T^*) = P_\rho.$$

Next $\rho^2 \in \sigma(w(T)w(T)^*)$ if and only if $0 \in \sigma(\phi_\rho(T, T^*))$. Further $\rho^2 \in \sigma_d(w(T)w(T)^*)$ if and only if $0 \in \sigma_d(\phi_\rho(T, T^*))$, and similarly for σ_e.

We now will prove some preliminary lemmas that we will need in order to make some reductions in our computation of $\sigma_d(w(T)w(T)^*)$. Our first result is Lemma 2.1.

LEMMA 2.1. *$T = V + F$ where V is unitary, and the rank of F is finite. Moreover $\sigma_e(V) = \sigma_e(T) = \sigma(T) \cap \partial D$.*

Proof. Set

$$V := T\left(I - \frac{1}{\|\mu\|^2}\mu \otimes \mu\right) + \frac{1}{\|\mu\|^2}\mu_* \otimes \mu$$

where

$$(a \otimes b)c := \langle c, b\rangle a$$

for $a, b, c \in H := H^2 \ominus mH^2$, $\mu(\zeta) := \bar{\zeta}(m(\zeta) - m(0))$, and $\mu_*(\zeta) := 1 - m(\zeta)\overline{m(0)}$. Then it is easy to check that V is unitary, rank $F \leqq 2$, and $\sigma_e(V) = \sigma_e(T) = \sigma(T) \cap \partial D$ (since V and T differ by a finite rank perturbation). □

Since we have assumed that w is not a constant times a Blaschke product, we see that

$$\phi_\rho | \partial D \not\equiv 0. \tag{3}$$

Set $\hat{\phi}_{0\rho}(\zeta) := \phi_\rho(\zeta, \bar{\zeta})$ for $\zeta \in \partial D$ (the unit circle). Then we have Corollary 2.2.

COROLLARY 2.2. *$0 \in \sigma_e(P_\rho)$ if and only if $\{\zeta \in \partial D: \hat{\phi}_{0\rho}(\zeta) = 0\} \cap \sigma(T) \neq \varnothing$.*

Proof. From (2.1) we get that $P_\rho = \phi_\rho(T, T^*) = \hat{\phi}_{0\rho}(V) + Q$, where Q is a finite rank operator. Thus from the fact that V is unitary and our above discussion, we see

$$\sigma_e(P_\rho) = \sigma_e(\hat{\phi}_{0\rho}(V)) = \hat{\phi}_{0\rho}(\sigma_e(V)) = \{\hat{\phi}_{0\rho}(\zeta): \zeta \in \sigma(T) \cap \partial D\},$$

which immediately implies our result. □

Remark 2.3. (i) Corollary 2.2 implies that in order to determine if $\rho^2 \in \sigma_d(w(T)w(T)^*)$ we can always assume that

$$\{\zeta \in \partial D: \hat{\phi}_{0\rho}(\zeta) = 0\} \cap \sigma(T) = \varnothing. \tag{4}$$

(ii) Let $\rho_{\text{ess}} := \rho_{\text{ess}}(w(T))$ denote the *essential norm* of $w(T)$ (i.e., the distance of $w(T)$ to the space of compact operators on H). Then we can show that [6], [9]

$$\begin{aligned}\rho_{\text{ess}}^2 &= \sup\{|w(\lambda)|^2: \lambda \in \sigma_e(T)\} \\ &= \sup\{|w(\lambda)|^2: \lambda \text{ is a singular point of } m \text{ on } \partial D\}.\end{aligned}$$

Notice that if $\rho > \rho_{\text{ess}}$, then automatically the assumption (4) given in (i) is satisfied. Moreover, the points in the set $\sigma(w(T)w(T)^*) \cap (\rho_{\text{ess}}^2, \infty)$ are precisely squares of the singular values of the operator $w(T)$. These points are part of the discrete spectrum of $w(T)w(T)^*$.

In summary, we have shown in this section that the computation of the discrete spectrum of $w(T)w(T)^*$ amounts to determining whether zero is an eigenvalue of finite multiplicity of the operator $\phi_\rho(T, T^*) \in \mathbf{C}[T, T^*]$ for given $\rho \in R$, where ϕ_ρ enjoys the properties (1)–(4). This is precisely the problem that we solve in the next section.

3. Main results. In this section we will formulate and prove our theorem on the computation of the discrete spectrum of operators of the form $w(T)w(T)^*$. From our discussion in § 2 we are reduced to the following kind of operator theoretic problem.

Let

$$\phi(z, \bar{z}) = \sum_{k,j=0}^{n} C_{kj} z^k \bar{z}^j \qquad (z \in \mathbf{C})$$

be a polynomial with the following properties:

(i) $\phi(z, \bar{z}) = \overline{\phi(z, \bar{z})}$, i.e., $C_{kj} = \bar{C}_{jk}$ $(0 \leqq j, k \leqq n)$.

(ii) $\phi | \partial D \not\equiv 0$. Set $\hat{\phi}_0(\zeta) := \phi(\zeta, \bar{\zeta})$, $\zeta \in \partial D$.

(iii) $\{\zeta \in \partial D: \hat{\phi}_0(\zeta) = 0\} \cap \sigma(T) = \varnothing$.

Now set

$$A := \phi(T, T^*) = \sum_{k,j=0}^{n} C_{kj} T^k T^{*j}.$$

From our arguments in § 2, we need to find a computable procedure for determining whether $0 \in \sigma_d(A)$. This will be done via a determinantal formula given in (3.6) and (3.8). For convenience, we now state the following reformulation of (2.2).

LEMMA 3.1. $0 \notin \sigma_e(A)$. *Equivalently* $0 \in \sigma_d(A)$ *if and only if* $0 \in \sigma(A)$.

Proof. This follows immediately from (2.2) and property (iii) above. □

Now in order to give our determinantal formula we will first have to compute the action of A on an element $g \in H := H^2 \ominus mH^2$. Accordingly, let

$$g = g_0 + g_1\zeta + \cdots \qquad (\zeta \in \partial D),$$
$$\bar{m}g = g_{-1}\bar{\zeta} + g_{-2}\bar{\zeta}^2 + \cdots.$$

Then

$$T^j g = P(\zeta^j g) = \zeta^j g - m(\zeta^{j-1} g_{-1} + \cdots + g_{-j}),$$
$$T^{*j} g = \bar{\zeta}^j g - (\bar{\zeta}^j g_0 + \cdots + \bar{\zeta} g_{j-1})$$

for $j \geqq 1$, and where $P: L^2 \to H$ denotes orthogonal projection.

Consequently,

$$\begin{aligned} Ag &= \sum_{k,j=0}^{n} C_{kj} P(\zeta^k T^{*j} g) \\ &= \sum_{k,j=0}^{n} C_{kj} P(\zeta^{k-j} g) - \sum_{j>0} C_{kj} \sum_{l=0}^{j-1} g_l P\zeta^{k-j+l} \\ &= \hat{\phi}_0(\zeta) g - \sum_{k>j} C_{kj} m \sum_{l=1}^{k-j} g_{-l} \zeta^{k-j-l} \\ &\quad - \sum_{k<j} C_{kj} \sum_{l=0}^{j-k-1} g_l \bar{\zeta}^{j-k-l} - \sum_{\substack{j>0 \\ k-j+l>0 \\ j>l}} C_{kj} g_l P\zeta^{k-j+l} \\ &= \hat{\phi}_0(\zeta) g - \sum_{l=1}^{n} g_{-l} m \sum_{k-j \geqq l} C_{kj} \zeta^{k-j-l} \\ &\quad - \sum_{l=0}^{n-1} g_l \sum_{j-k>l} C_{kj} \bar{\zeta}^{j-k-l} - \sum_{l=0}^{n-1} g_l \sum_{k+l \geqq j>l} C_{kj} P\zeta^{k+l-j}. \end{aligned}$$

Set

$$\phi_l^+(\zeta) := \sum_{k-j \geqq l} C_{kj} \zeta^{k-j-l} \qquad (1 \leqq l \leqq n), \tag{5}$$

$$\phi_l^-(\zeta) := \sum_{j-k>l} C_{kj} \bar{\zeta}^{j-k-l} \qquad (0 \leqq l \leqq n-1), \tag{6}$$

$$\begin{aligned} \phi_l(\zeta) &:= \sum_{k+l \geqq j>l} C_{kj} P\zeta^{k+l-j} \qquad (0 \leqq l \leqq n-1) \\ &= \sum_{k+l \geqq j>l} C_{kj} (P(\zeta^{k+l-j} P1))(\zeta) \qquad (7) \\ &= \sum_{k+l \geqq j>l} C_{kj} [\zeta^{k+l-j} \mu_*(\zeta) - m(\zeta^{k+l-j-1} \mu_{*,-1} + \cdots + \mu_{*,-(k+l)+j})] \end{aligned}$$

where

$$\mu_{*,-j} := \bar{m}_j \quad \text{for } j \geqq 1. \tag{8}$$

(We are setting $m(\zeta) = \sum_{j=0}^{\infty} m_j \zeta^j$.)

We can now summarize our above computation by Lemma 3.2.

LEMMA 3.2. *We have for $g \in H$ that*

$$Ag = \hat{\phi}_0 g - \sum_{l=1}^{n} g_{-l} m \phi_l^+ - \sum_{l=0}^{n-1} g_l \phi_l^- - \sum_{l=0}^{n-1} g_l \phi_l \tag{9}$$

where $\hat{\phi}_0$, ϕ_l^-, ϕ_l^+, and ϕ_l are explicitly given from (ii), (5), (6), (7), *respectively, above.*

Remark 3.3. Notice from (3.1) that $0 \in \sigma(A)$ if and only if there exists $g \in H$, $g \neq 0$ such that $Ag = 0$. From (3.2) we can compute the action of A on g. We assume from now on that $Ag = 0$. Then by (3.2) (for $Ag = 0$), we have that

$$\hat{\phi}_0 g = \sum_{l=1}^{n} g_{-l} m \phi_l^+ + \sum_{l=0}^{n-1} g_l (\phi_l^- + \phi_l). \tag{10}$$

Now define

$$\psi(z) := \sum_{k,j=0}^{n} C_{kj} z^{n+k-j} \qquad (z \in \mathbf{C}). \tag{11}$$

Then multiplying (10) by ζ^n, we can easily deduce that

$$\psi(z) g(z) = \sum_{l=1}^{n} g_{-l} m(z) \phi_l^+(z) z^n + \sum_{l=0}^{n-1} g_l (\psi_l^-(z) + z^n \phi_l(z)) \tag{12}$$

for $z = \zeta \in \partial D$, where

$$\psi_l^-(z) := \sum_{j-k>l} C_{kj} z^{n+l+k-j} \tag{13}$$

for $z \in \mathbf{C}$.

We now make a technical assumption in order to simplify our exposition. This assumption of genericity will be *removed* when we state our final result in (3.8).

Assumption 3.4. All the zeros of ψ are distinct and different from zero.

We now come to the following result.

LEMMA 3.5. *Under assumption* (3.4), *there exist* $z_1, z_2, \cdots, z_p \in D$, $\zeta_1, \zeta_2, \cdots, \zeta_q \in \partial D \backslash \sigma(T)$, $2p + q = 2n$, *such that* $\psi(z) = \alpha(z - z_1) \cdots (z - z_p)(z - 1/\bar{z}_1) \cdots (z - 1/\bar{z}_p)(z - \zeta_1) \cdots (z - \zeta_q)$ *for some* $\alpha \neq 0$.

Proof. From properties (i) and (ii) at the beginning of this section we have that

$$\begin{aligned} z^{2n} \overline{\psi(1/\bar{z})} &= z^{2n} \sum_{k,j=0}^{n} \overline{C_{kj} (1/\bar{z})^{n+k-j}} \\ &= \sum_{k,j=0}^{n} \bar{C}_{kj} z^{n-k+j} \\ &= \sum_{k,j=0}^{n} C_{jk} z^{n-k+j} = \psi(z). \end{aligned}$$

From (ii), $\psi \not\equiv 0$. Denote by $\zeta_1, \zeta_2, \cdots, \zeta_q$ the zeros of ψ on ∂D. From our above computation it follows that if $\psi(z_0) = 0$ for $z_0 \in D$ with $z_0 \neq 0$, then $\psi(1/\bar{z}_0) = 0$ also. This yields the representation of ψ. Finally $\zeta_j \notin \sigma(T)$ by (iii) for $j = 1, \cdots, q$. □

We are almost done! Indeed all the functions in (12) are analytic in a neighborhood of D except $\sigma(T)\cap\partial D$. This allows us to set $z=z_1, z_2, \cdots, z_p, \zeta_1, \cdots, \zeta_q$ in (12) obtaining

$$\sum_{l=1}^{n} g_{-l}m(z_r)\phi_l^+(z_r)z_r^n+\sum_{l=0}^{n-1} g_l(\psi_l^-(z_r)+z_r^n\phi_l(z_r))=0 \quad \text{for } 1\leqq r\leqq p, \tag{14}$$

$$\sum_{l=1}^{n} g_{-l}m(\zeta_s)\phi_l^+(\zeta_s)\zeta_s^n+\sum_{l=0}^{n-1} g_l(\psi_l^-(\zeta_s)+\zeta_s^n\phi_l(\zeta_s))=0 \quad \text{for } 1\leqq s\leqq q. \tag{15}$$

Now multiplying (10) by $\overline{\zeta^n m(\zeta)}$, we see

$$[\bar{z}^n\hat{\phi}_0](z)(\bar{m}g)(z)=\sum_{l=1}^{n} g_{-l}[\bar{z}^n\phi_l^+](z)+\sum_{l=0}^{n-1} g_l\{[\bar{z}^n\bar{m}\phi_l^-](z)+[\bar{z}^n\bar{m}\phi_l](z)\} \tag{16}$$

where $z=\zeta\in\partial D$ and all the functions are analytic in $\bar{z}$. Note that even though this equation has been derived on ∂D, it is valid on the complement of D if we replace $\bar{z}$ by $1/z$ for $|z|>1$.

Now set

$$\psi_l^+(z):=(\bar{z}^n\phi_l^+)(z)=\sum_{k-j\geqq l} C_{kj}\bar{z}^{n+j-k+l} \qquad (1\leqq l\leqq n) \tag{17}$$

and

$$\begin{aligned}\psi_l(z)&:=\bar{z}^n(\bar{m}\phi_l)(z) \qquad (0\leqq l\leqq n-1)\\ &=\sum_{k+l\geqq j>l} C_{kj}\bar{z}^{n+j-k-l}[(\overline{m(z)}-\overline{m(0)})-(\bar{z}^{n-k-l+j+1}\mu_{*,-1}\\ &\qquad +\cdots+\bar{z}^n\mu_{*,-(k+l)+j})]\end{aligned} \tag{18}$$

for $z=\zeta\in\partial D$. Once again $\psi_l^+(z)$, $\psi_l(z)$ admit analytic extensions to the complement of $\bar{D}$ if we replace $\bar{z}$ by $1/z$ for $|z|>1$.

Moreover for $1\leqq r\leqq p$,

$$\bar{z}_r^n\hat{\phi}_0\left(\frac{1}{\bar{z}_r}\right)=\bar{z}_r^{2n}\psi\left(\frac{1}{\bar{z}_r}\right)=0. \tag{19}$$

(Notice we are setting $\hat{\phi}_0(z):=\phi(z,1/z)$ for $z\in\mathbf{C}$.) Then from (16) we see that for $1\leqq r\leqq p$,

$$\sum_{l=1}^{n} g_{-l}\psi_l^+\left(\frac{1}{\bar{z}_r}\right)+\sum_{l=0}^{n-1} g_l\left\{z^{-n}\bar{m}\phi_l\left(\frac{1}{\bar{z}_r}\right)+\psi_l\left(\frac{1}{\bar{z}_r}\right)\right\}=0. \tag{20}$$

(We are setting $\bar{m}(z):=\overline{m(1/\bar{z})}$ for $|z|>1$.)

Finally we note that if $g_{-n}=\cdots=g_{n-1}=0$, then from (3.2) we have that $\phi_0 g=0$ (note we have taken g such that $Ag=0$), which by property (ii) above implies that $g=0$.

We now come to the final point in our computations. Namely the above argument shows that $0\in\sigma(A)$ if and only if the characteristic determinant of the $2n$ equations in the $2n$ "unknowns" $g_{-n}, \cdots, g_{n-1}$ is zero. We can write this determinant quite explicitly. Indeed in order to do this, let us introduce the notation

$$M(\theta_1, \cdots, \theta_R; \xi_1, \cdots, \xi_N)=\begin{bmatrix}\theta_1(\xi_1) & \cdots & \theta_R(\xi_1)\\ \vdots & & \vdots\\ \theta_1(\xi_N) & \cdots & \theta_R(\xi_N)\end{bmatrix} \tag{21}$$

for functions $\theta_1, \cdots, \theta_R$ well defined in a neighborhood of $\xi_1, \cdots, \xi_N$ with the ξ_i distinct for $i = 1, \cdots, N$.

Using this notation, our preceding arguments prove the following theorem.

THEOREM 3.6. *Under Assumption* 3.4, $0 \in \sigma(A)$ *if and only if*

$$\det \begin{bmatrix} M^- & M^+ \\ N^- & N^+ \\ M_*^- & M_*^+ \end{bmatrix} = 0 \tag{22}$$

where

$$M^- := M(z^n m\phi_n^+, \cdots, z^n m\phi_1^+; z_1, \cdots, z_p), \tag{23}$$

$$M^+ := M(\psi_0^- + z^n\phi_0, \cdots, \psi_{n-1}^- + z^n\phi_{n-1}; z_1, \cdots, z_p), \tag{24}$$

$$M_*^- := M(\psi_n^+, \cdots, \psi_1^+; 1/\bar{z}_1, \cdots, 1/\bar{z}_p), \tag{25}$$

$$M_*^+ := M(\psi_0 + z^{-n}\bar{m}\phi_0^-, \cdots, \psi_{n-1} + z^{-n}\bar{m}\phi_{n-1}^-; 1/\bar{z}_1, \cdots, 1/\bar{z}_p). \tag{26}$$

N^- *and* N^+ *are defined as in* (23) *and* (24) *by replacing* $z_1, \cdots, z_p$ *with* $\zeta_1, \cdots, \zeta_q$.

Proof. Write the characteristic determinant of the system of equations in (14), (15), (20)! □

Remark 3.7. We will now eliminate Assumption 3.4 in Theorem 3.6. Note that if the roots of ψ are not distinct perturbing ψ by ε, ε a suitable sufficiently small number, will assure that the corresponding ψ_ε does have distinct roots different from zero.

Before stating our result, we will need to extend the definition of M in (21) to the case where the ξ_i have multiplicities. Indeed, we set

$$M(\theta_1, \cdots, \theta_R; \xi_1, \cdots, \xi_1, \cdots, \xi_S, \cdots, \xi_S) := \begin{bmatrix} \theta_1(\xi_1) & \theta_2(\xi_1) & \cdots & \theta_R(\xi_1) \\ \theta_1'(\xi_1) & \theta_2'(\xi_1) & \cdots & \theta_R'(\xi_1) \\ \vdots & \vdots & & \vdots \\ \theta_1^{(N_1-1)}(\xi_1) & \theta_2^{(N_1-1)}(\xi_1) & \cdots & \theta_R^{(N_1-1)}(\xi_1) \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ \theta_1(\xi_S) & \theta_2(\xi_S) & \cdots & \theta_R(\xi_S) \\ \vdots & \vdots & & \vdots \\ \theta_1^{(N_S-1)}(\xi_S) & \theta_2^{(N_S-1)}(\xi_S) & \cdots & \theta_R^{(N_S-1)}(\xi_S) \end{bmatrix} \tag{27}$$

where ξ_i has multiplicity N_i for $i = 1, \cdots, S$ in M, and the functions $\theta_1, \cdots, \theta_R$ are analytic in a neighborhood of $\xi_1, \cdots, \xi_S$. (For θ analytic, $\theta^{(N)}$ denotes the derivative of order N.)

Then taking $\varepsilon \to 0$ in our above argument, we easily get the following corollary.

COROLLARY 3.8. *In complete generality* (*i.e., without Assumption* 3.4), *we have that* $0 \in \sigma(A)$ *if and only if the determinant* (22) *is zero where we use the definition* (27) *of* M *in* (23)–(26) *above, and each root of* $\psi(z)$ *is counted according to its multiplicity.*

Remark 3.9. (i) The determinantal formula (22) gives us an explicit expression for determining the invertibility of the operator $A \in \mathbf{C}[T, T^*]$. Moreover from our discussion in § 2, we can now also find $\sigma_d(w(T)w(T)^*)$. We will apply this to an example in § 4.

(ii) We should also note that the multiplicity of zero as a root of $\det M$ (in (22)) is its multiplicity as an eigenvalue of A. Moreover, from (10) we have a formula for the corresponding *eigenvectors.* Hence we have a general procedure for computing the multiplicity of the singular values of $w(T)$ and for the corresponding Schmidt vectors.

4. Example. In this section we illustrate via an example some of the computational issues involved in the determinantal scheme for computing the singular values of the operators $w(T)$ that we have discussed above. We are convinced that it will be possible in the near future to implement on a computer the formulae discussed in § 3, so that hopefully these ideas can become of practical use for some applied problems.

Let $w(z)=z^2+1$, and let $m\in H^\infty$ be a nonconstant inner function. We want to study the singular values of the corresponding operator $w(T)$. First note that $\|w\|_\infty=2$, and w attains its maximum at ± 1. If $\pm 1\in\sigma_e(T)$ (i.e., if m is singular at ± 1), then all of the s-numbers of $w(T)$ will be equal to two. Consequently, we will assume for now on that $\pm 1\notin\sigma_e(T)$, and $\rho_{\text{ess}}:=\rho_{\text{ess}}(w(T))<2$. We will study the invertibility of P_ρ (notation as § 2) for ρ contained in the interval $(\rho_{\text{ess}},2)$.

The computation of the determinantal formula for the singular values of $w(T)$ in $(\rho_{\text{ess}},2)$ is now quite elementary following the arguments of § 3. Indeed using the notation of § 2 (see (11)), we have that

$$\psi_\rho(z)=-\frac{1}{\rho^2}z^4-\left(\frac{2}{\rho^2}-1\right)z^2-\frac{1}{\rho^2}.$$

We can calculate that in the interval of interest all of the roots of $\psi_\rho(z)$ lie on ∂D and are distinct. The exact formulae for these roots are $\zeta_1=e^{i\theta/2}$, $\xi_2=-e^{i\theta/2}$, $\zeta_3=e^{-i\theta/2}$, $\zeta_4=-e^{-i\theta/2}$, where

$$\theta:=\arctan\frac{\rho\sqrt{1-\rho^2/4}}{(1-\rho^2/2)}\qquad(0<\theta<\pi/2).$$

(Notice that $\zeta_3=\bar{\zeta}_1$, $\zeta_4=\bar{\zeta}_2$.)

If we now follow the recipe of § 3, we see (the computations actually were quite easy!) that the singular values of $w(T)$ in the interval $(\rho_{\text{ess}},2)$ may be derived from the determinant of the following 4×4 matrix:

$$M=[N^-\quad N^+] \tag{28}$$

where

$$N^-:=\begin{bmatrix}-(1/\rho^2)\zeta_1^2m(\zeta_1) & -(1/\rho^2)\zeta_1^3m(\zeta_1)\\ \vdots & \vdots\\ -(1/\rho^2)\zeta_4^2m(\zeta_4) & -(1/\rho^2)\zeta_4^3m(\zeta_4)\end{bmatrix},$$

$$N^+=\begin{bmatrix}-(1/\rho^2)(1+\zeta_1^2\mu_*(\zeta_1)) & -(1/\rho^2)(\zeta_1+\zeta_1^2(\zeta_1\mu_*(\zeta_1)-m(\zeta_1)\bar{m}_1))\\ \vdots & \vdots\\ -(1/\rho^2)(1+\zeta_4^2\mu_*(\zeta_4)) & -(1/\rho^2)(\zeta_4+\zeta_4^2(\zeta_4\mu_*(\zeta_4)-m(\zeta_4)\bar{m}_1))\end{bmatrix}.$$

(Recall that $\mu_*(\zeta)=1-m(\zeta)\overline{m(0)}$, and $m_1:=dm/d\zeta|_{\zeta=0}$.)

Notice that in this formula m appears as a "parameter," and that, in general, the size and complexity of the matrix given in (22) above only depends on the "weighting" function w. We checked (28) for the trivial case of $m(z)=z$, and got (of course) the obvious answer that the unique root of $\det M$ in $(0,2)$ is 1.

In conclusion, the determinantal formula (22) offers a very general theoretical procedure for the computation of the discrete spectrum of operators of the form $w(T)w(T)^*$ for both rational and irrational inner functions $m\in H^\infty$. The writing of appropriate software for actually carrying this out should make a very interesting project.

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem*, Math. USSR-Sb., 15 (1971), pp. 31-73.

[2] C. FOIAS AND A. TANNENBAUM, *On the Nehari problem for a certain class of L^∞-functions appearing in control theory*, J. Funct. Anal., 74 (1987), pp. 146-159.

[3] C. FOIAS, A. TANNENBAUM, AND G. ZAMES, *On the H^∞-optimal sensitivity problem for systems with delays*, SIAM J. Control Optim., 25 (1987), pp. 686-706.

[4] ———, *Sensitivity minimization for arbitrary* SISO *distributed plants*, Systems Control Lett., 8 (1987), pp. 189-195.

[5] B. A. FRANCIS, *A Course in H^∞ Control Theory*, Lecture Notes in Control and Information Sciences 88, Springer-Verlag, New York, 1987.

[6] N. K. NIKOLSKII, *A Treatise on the Shift Operator*, Springer-Verlag, New York, 1986.

[7] S. C. POWER, *Hankel Operators on Hilbert Space*, Pitman, Boston, MA, 1982.

[8] D. SARASON, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179-203.

[9] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[10] G. ZAMES, A. TANNENBAUM, AND C. FOIAS, *Optimal interpolation in H^∞: a new approach*, in Proc. of the Conference on Decision and Control, December 1986, pp. 350-355.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

005

A FUNDAMENTAL INTEGRAL RELATION OF SCATTERING THEORY*

MARGARET CHENEY†, JAMES H. ROSE‡, AND BRIAN DEFACIO§

Abstract. This paper concerns three-dimensional scattering and inverse scattering for a variety of time-reduced wave equations. The main result is an integral relation that relates the wavefield to the scattering data. A rigorous derivation of this integral relation is given for a specific class of linear, scalar wave equations. Three specific examples of wave equations to which these results apply are then considered: the Schrödinger equation with complex potential, the wave equation with variable speed, and the acoustic equation with variable density and speed. Finally, three consequences of the integral relation are considered. First, this equation is used to derive a generalized optical theorem for the above-mentioned class of wave equations. Second, the relation's implications for long wavelength scattering are discussed. Third, this equation is shown to lie at the heart of certain inverse scattering methods.

Key words. scattering theory, inverse problems, optical theorem, Marchenko equation

AMS(MOS) subject classification. 35P25

1. Introduction. In 1980, R. G. Newton published two exact three-dimensional inverse scattering methods for Schrödinger's equation [1]. Recently [2], one of these methods has been generalized to apply to scattering from an inhomogeneous medium. This may be an important step towards solving the multidimensional inverse problem for acoustic, elastic, and electromagnetic scattering.

The method mentioned above proceeds in two steps. First a linear integral equation is used to compute the wavefield everywhere from the scattering data. Then the properties of the scatterer are inferred from the reconstructed wavefield. The frequency domain version of the integral equation just mentioned is the subject of this paper.

These frequency domain integral relations relating the wavefield to the scattering data have been developed in a number of contexts. For obstacle scattering, such a relation was found by Lax and Phillips [3]. For quantum scattering, a similar relation appears in Schmidt [4] and in Newton [5]. A very similar integral relation was found by the authors [6] for scattering governed by a hyperbolic equation closely related to the Schrödinger equation. The authors found [2] that essentially the same integral relation holds for scattering of waves in inhomogeneous media.

The fact that all these integral relations are identical is rather surprising. It suggests that they reflect something common to scattering problems. Indeed, in a recent letter [7], the authors have derived this integral relation on physical grounds for a wide class of scattering experiments. Scattering experiments in this class were assumed to be governed by a variety of linear hyperbolic wave equations, but exact conditions on the equations were not stated.

In this paper, we give a careful mathematical derivation of the integral relation for a specific class of three-dimensional scattering experiments. This derivation has a number of advantages over the derivations of [4], [5], and [6]. First, the proofs of [4] and [5] are worked out only for the Schrödinger equation, whereas our proof holds

* Received by the editors October 12, 1986; accepted for publication (in revised form) October 10, 1987.

† Department of Mathematics, Duke University, Durham, North Carolina 27706. The work of this author was supported by Office of Naval Research contract N00014-85-K-0224.

‡ Center for Nondestructive Evaluation, Iowa State University, Ames, Iowa 50011. The work of this author was supported by the National Science Foundation University/Industry Center for Nondestructive Evaluation at Iowa State University.

§ Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211.

for a wide class of wave equations. It is especially hard to see how to generalize the argument of [5] to other equations because that argument depends heavily on the use of spectral theory. Many interesting equations cannot be considered as eigenvalue equations in a natural way. The earlier work of the authors [2] is applicable to a fairly wide class of equations, but the proof in [2] required smoothness assumptions on the scatterer. This is quite a drawback, because many of the applications of inverse scattering deal with imaging discontinuities (e.g., an object imbedded in a solid). The proof in this paper is valid for discontinuous scatterers.

The various proofs in this paper are carried out in the frequency domain. The Fourier transform of the integral relation to obtain a time-domain equation will be only briefly mentioned. Our use of the frequency domain underlines a continuing tension in the development of inverse scattering theory. The theory is most transparent in the time domain where causality appears naturally. However, the mathematical development is simpler in the frequency domain where considerably more is known.

The structure of this paper is as follows. In § 2, the problem is discussed and the physical picture is explained. Section 3 contains a derivation of the integral relation mentioned above. In § 4, several wave equations which arise from different classes of physical problems are discussed as examples. In § 5, the integral relation is used to obtain various results. First we show that it leads to a straightforward proof of a generalized optical theorem for these wave equations. Then we show that it leads to certain results in the long wavelength scattering limit. In particular, we show that the long wavelength phase for acoustic wave scattering satisfies certain symmetry conditions. Finally, § 6 contains applications of the integral relation to inverse scattering problems.

2. Physical picture and statement of results. We assume that the scattering is governed by the equation

$$(\nabla^2 + k^2)\psi = V\psi. \tag{2.1}$$

Here ∇^2 is the Laplacian on R^n, k is a real scalar, and V is a linear operator satisfying certain hypotheses below.

In defining scattering solutions of (2.1), it will be useful to consider the operator $(\nabla^2 + k^2)$. For the time being, we take its domain to be H^2, the space of functions with derivatives up to order two in L^2. Although we will ultimately be interested in real values of k, it is helpful to consider $\nabla^2 + k^2$ also for complex k. For nonreal k, $\nabla^2 + k^2$ maps H^2 onto L^2 and has a bounded inverse [8]. We denote the inverse by $G_0 = (\nabla^2 + k^2)^{-1}$. For real k, G_0 is no longer bounded as a map from L^2 to H^2. However, $\nabla^2 + k^2$ still has fundamental solutions or Green's functions. We denote by $G_0^+ (G_0^-)$ the fundamental solution which in a certain sense (to be specified later) is the limit of the kernel of G_0 as k approaches the real axis from the upper (lower) half-plane [9]. Specifically for $n = 3$ we have

$$G_0^{\pm}(k, r) = -(4\pi r)^{-1} \exp(\pm ikr) \tag{2.2}$$

where r is the length of $\vec{x}$ for $\vec{x}$ in R^3. We will use (2.2) to obtain an integral equation for solutions of (2.1).

We are especially interested in scattering solutions of (2.1); thus we are interested in solutions which for large $\vec{x}$ behave like plane waves $\exp(ik\hat{e} \cdot \vec{x})$ propagating in direction $\hat{e}$, where $\hat{e}$ is a unit vector in $\mathbb{R}^n$. We therefore define the scattering solutions $\psi^{\pm}$ by the integral equation

$$\psi^{\pm}(k, \hat{e}, \vec{x}) = \exp(ik\hat{e} \cdot \vec{x}) + \int G_0^{\pm}(k, |\vec{x} - \vec{y}|)(V\psi^{\pm})(k, \hat{e}, \vec{y})\, d\vec{y}. \tag{2.3$\pm$}$$

(The spaces in which the solutions $\psi^{\pm}$ lie will be discussed later.) In what follows we will also use the symbols $G_0^{\pm}V$ to mean the integral operator of (2.3±). Moreover, we will assume below (and, in fact show for a certain class of V's) that (2.3±) has a unique solution $\psi^{\pm}$. Physically, this means that the system responds in a well-defined way to the incident plane wave.

In many cases, expanding (2.3±) for large $|\vec{x}|$ shows that the interaction of the incident plane wave with the scatterer gives rise to a spherically spreading wave. For example, for $n=3$, ψ^{+} can be written asymptotically as

$$\psi^{+}(k, \hat{e}, \vec{x}) = \exp(ik\hat{e}\cdot\vec{x}) + \frac{A(k, \hat{x}, \hat{e})}{|\vec{x}|} e^{ik|\vec{x}|} + O(|x|^{-2}). \tag{2.4}$$

Here $\hat{e}$ and $\hat{x}$ denote the direction of incidence and scattering, respectively, and A denotes the scattering amplitude. For $n=3$, the scattering amplitude is given by the formula [5]

$$A(k, \hat{e}, \hat{e}') = -\frac{1}{4\pi}\int \exp(-ik\hat{e}\cdot\vec{y})(V\psi^{+})(k, \hat{e}', \vec{y})\, d\vec{y}. \tag{2.5}$$

Sections 3 and 4 require only definition (2.5) of the scattering amplitude. Its relation (2.4) to the physical picture is irrelevant to the mathematics in these two sections. The goal of this paper is to prove, for $n=3$, the following relation between ψ^{+}, ψ^{-}, and A:

$$\psi^{+}(k, \hat{e}, \vec{x}) = \psi^{-}(k, \hat{e}, \vec{x}) + \frac{ik}{2\pi}\int_{S^2} A(k, \hat{e}', \hat{e})\psi^{-}(k, \hat{e}', \vec{x})\, d\hat{e}', \tag{2.6}$$

where S^2 denotes the unit sphere in $\mathbb{R}^3$. After proving (2.6), we will consider some of the conclusions that can be drawn from this equation.

3. Proof of (2.6). In this section, we prove (2.6) for scattering governed by an operator V which satisfies conditions (H1) and (H2) (uniqueness and decay, see below) for $n=3$. Since (H1) and (H2) are rather abstract, the question then arises, for what operators V do (H1) and (H2) hold? Theorem 3.1 gives a specific class of operators for which (H1) and (H2) hold. As we will see in the examples, this class includes several physically interesting wave equations.

We define the weighted L^2 space on R^n

$$L^{2,s} = \{u: (1+|\vec{x}|^2)^{s/2}u \in L^2\}$$

and the weighted Sobolev space $H^{m,s}$, which is the space of functions with derivatives up to order m in $L^{2,s}$.

THEOREM 3.1. *Suppose V is given by*

$$V(f) = \sum_{j=1}^{n} a_j(k, \vec{x})\frac{\partial}{\partial x_j}f(\vec{x}) + b(k, \vec{x})f(\vec{x}), \tag{3.1}$$

where the coefficients satisfy the following conditions:

(a) *For almost all $\vec{x}$, the a_j and b are entire functions of k.*

(b) *For each complex k, there is some $\varepsilon > 0$ such that for some α with $0 < \alpha < 4$,*

$$\sup_{\vec{x}\in\mathbb{R}^n}\left[\int_{|\vec{y}-\vec{x}|<1}(1+|\vec{y}|^2)^{n+\varepsilon}|b(k, \vec{y})|^2|\vec{y}-\vec{x}|^{\alpha-n}\, d\vec{y}\right]$$

is finite and

$$\int_{|\vec{y}-\vec{x}|<1}(1+|\vec{x}|^2)^{n+\varepsilon}|b(k, \vec{x})|^2\, d\vec{x} \to 0 \quad \text{as } |\vec{y}|\to\infty.$$

(*This hypothesis will be satisfied if $b(k, \vec{x})$ is locally in L^2 and is $O(|\vec{x}|^{-n-2\varepsilon})$ as $|\vec{x}|\to\infty$.*)

(c) *For each* $j=1, 2, \cdots, n$ *and each complex* k, *there is some* β *with* $0<\beta<2$ *such that for the above* ε,

$$\sup_{\vec{x}\in\mathbb{R}^n} \int_{|\vec{y}-\vec{x}|<1} (1+|\vec{y}|^2)^{n+\varepsilon} |a_j(k, \vec{y})|^2 |\vec{y}-\vec{x}|^{\beta-n}\, d\vec{y}$$

is finite and

$$\int_{|\vec{y}-\vec{x}|<1} (1+|\vec{y}|^2)^{n+\varepsilon} |a_j(k, \vec{y})|^2\, d\vec{y} \to 0 \quad \text{as } |\vec{x}|\to\infty.$$

(*This hypothesis will be satisfied if* $a_j(k, \vec{x})$ *is bounded and is* $O(|\vec{x}|^{-n-2\varepsilon})$ *as* $|\vec{x}|\to\infty$.)

(d) *The operator* $\nabla^2 - V + k^2 : H^2 \to L^2$ *is invertible for some* k *in the set* $\{k: \operatorname{Im} k>0, k\notin[0,\infty)\}$.

Then for almost all real k, V *satisfies the following*:

(H1+) (*uniqueness*) $I - G_0^+ V$ *is invertible on* $H^{2,-s}$ *for* $s = n/2+\varepsilon/2$.

(H2) (*decay*) *For some* $\varepsilon>0$, V *maps* $H^{2,-s}$ *into* $L^{2,s}$ *for* $s = n/2+\varepsilon/2$.

A similar statement holds if $\operatorname{Im} k>0$ *in* (d) *is replaced by* $\operatorname{Im} k<0$; *the conclusion* (H1+) *is then replaced by*

(H1−) $I - G_0^- V$ *is invertible on* $H^{2,-s}$ *for* $s = n/2+\varepsilon/2$.

Remark. As we will see in the examples, it is often easy to show that hypothesis (d) is satisfied. It does not appear to be known whether hypotheses (b) and (c) always imply (d); there are examples of differential operators with no resolvent set (see [10]) but they are not of the form considered here.

Proof. We will show that the operator $G_0 V$ is compact on $H^{2,-s}$ for $s = n/2+\varepsilon/2$. This fact plus condition (a) allows us to apply the analytic Fredholm Theorem [10], [11] to $I - G_0 V$ in the upper-half k-plane. The analytic Fredholm Theorem implies that one of two things must be true about $I - G_0 V$; either

(i) $I - G_0 V$ is invertible nowhere in $\operatorname{Im} k \geqq 0$, or

(ii) $I - G_0 V$ is invertible in $\operatorname{Im} k \geqq 0$, except possibly on a discrete set in $\operatorname{Im} k>0$ whose limit points on the real axis are of Lebesgue measure zero. Thus $I - G_0^+ V$ is invertible for almost all real k.

We will use hypothesis (d) to rule out alternative (i).

We will prove compactness by the following argument. First, Agmon [9] has shown that for $\operatorname{Im} k \geqq 0$, G_0 is a bounded operator mapping $L^{2,s}$ into $H^{2,-s}$ for $s>\frac{1}{2}$. We need only show that $V: H^{2,-s} \to L^{2,s}$ is compact. Thus we will prove (H2) in the process of proving (H1+).

First we consider the last term of (3.1). Condition (b) is precisely the condition guaranteeing that the operator of multiplication by $(1+|x|^2)^{n+\varepsilon/2} b(k, x)$ is compact as a mapping from H^2 into L^2 [12]. In other words, b is compact as a mapping from H^2 to $L^{2,n+\varepsilon}$. Moreover, by the following argument, b is compact from $H^{2,s}$ to $L^{2,s+n+\varepsilon}$ for any s. The operator of multiplication by $(1+|x|^2)^{s/2}$ is bounded from $H^{2,r}$ to $H^{2,r-s}$ and from $L^{2,r}$ to $L^{2,r-s}$ for any r. Multiplication by b can thus be considered as the composition

$$H^{2,s} \xrightarrow[(1+|x|^2)^{s/2}]{} H^2 \underset{b}{\to} L^{2,n+\varepsilon} \xrightarrow[(1+|x|^2)^{-s/2}]{} L^{2,s+n+\varepsilon}.$$

Since the composition of bounded operators and compact operators is again compact, this shows that b is compact as a mapping from $H^{2,s}$ to $L^{2,s+n+\varepsilon}$. For $s = n/2+\varepsilon/2$, $b: H^{2,-s} \to L^{2,s}$ is compact.

A very similar argument works for the other terms in (3.1). We consider instead the composition

$$H^{2,s} \xrightarrow[\partial/\partial x_j]{} H^{1,s} \xrightarrow[(1+|x|^2)^{s/2}]{} H^1 \xrightarrow[a_j]{} L^{2,n+\varepsilon} \xrightarrow[(1+|x|^2)^{-s/2}]{} L^{2,s+n+\varepsilon}.$$

Again condition (c) is precisely the right condition to ensure that the map $a_j: H^1 \to L^{2,n+\varepsilon}$ is compact [12].

We have now shown that G_0V is compact on $H^{2,-s}$ for $s = n/2 + \varepsilon/2$. Now we must rule out alternative (i) of the analytic Fredholm Theorem. For this we use hypothesis (d). Note that (d) is a statement about invertibility on L^2, not on $H^{2,-s}$. We show now that for the k in hypothesis (d), $I - G_0V$ is invertible on $H^{2,-s}$. We do this by an argument similar to that in [9]. We will use the identity

$$(\nabla^2 + k^2)^{-1} = (\nabla^2 - V + k^2)^{-1} - (\nabla^2 + k^2)^{-1} V (\nabla^2 - V + k^2)^{-1}$$

which we rewrite using $G_0 = (\nabla^2 + k^2)^{-1}$ and $G = (\nabla^2 - V + k^2)^{-1}$ as $G_0 = G - G_0VG$ or

$$G_0 = (I - G_0V)G. \tag{3.2}$$

This identity is valid for values of k for which both $(\nabla^2 + k^2)$ and $\nabla^2 - V + k^2$ are invertible as maps from H^2 to L^2. In particular, hypothesis (d) asserts that there is at least one k for which (3.2) holds. At this k, the similar identity $G_0 = G - GVG_0$ or

$$G_0 = G(I - VG_0) \tag{3.3}$$

also holds. Since $I - VG_0$ maps L^2 into L^2 and since the range of G_0 is all of H^2, (3.3) shows that G maps L^2 onto H^2. Equation (3.2) then shows that $I - G_0V$ maps H^2 onto H^2. Therefore $(I - G_0V)H^{2,-s}$ must contain all of H^2. However, since G_0V is compact, the range of $I - G_0V$ in $H^{2,-s}$ is closed. Since H^2 is dense in the range of $I - G_0V$ (in the $H^{2,-s}$ norm), it must be true that the range of $I - G_0V$ is all of $H^{2,-s}$. $I - G_0V$ is therefore invertible on $H^{2,-s}$. □

The condition (H1±) is useful for working with (2.3±). In particular, condition (H1±) means that for f in $H^{2,-s}$, the integral equation

$$h = f + \int G_0^{\pm} Vh$$

has a unique solution h in $H^{2,-s}$. In particular, equations (2.3±) each have unique solutions in $H^{2,-s}$. The scattering solutions $\psi^{\pm}$ are therefore well defined. Since ψ^+ is used in the definition (2.5) of the scattering amplitude A, condition (H1+) should be satisfied whenever A is used.

We note that V is assumed to be independent of the boundary conditions of the problem in § 2; in particular it is independent of the variable $\hat{e}$.

LEMMA 3.2. *Suppose conditions* (H1+) *and* (H2) *hold for* $n = 3$. *Then* A *defined by* (2.5) *is bounded.*

Proof. Write the integrand of (2.5) as the product $\exp(-ik\hat{e}\cdot\vec{y})(1+|\vec{y}|^2)^{-s/2}$ times $(1+|\vec{y}|^2)^{s/2}(V\psi^+)(k, \hat{e}', \vec{y})$ for $s = 3/2 + \varepsilon/2$ and apply the Schwarz inequality. By condition (H2), $(1+|\vec{y}|^2)^{s/2}V\psi^+$ has finite L^2 norm. □

PROPOSITION 3.3. *Suppose conditions* (H1+) *and* (H2) *hold for* $n = 3$. *Then*

$$\int_{S^2} A(k, \hat{e}, \hat{e}') \exp(ik\hat{e}\cdot\vec{x})\, d\hat{e}$$
$$= \frac{2\pi}{ik}\left[\psi^+(k, \hat{e}', \vec{x}) - \exp(ik\hat{e}'\cdot\vec{x}) - \int G_0^-(k, |\vec{x}-\vec{y}|)(V\psi^+)(k, \hat{e}', \vec{y})\, d\vec{y}\right]. \tag{3.4}$$

Proof. We use (2.5) in the left side of (3.4). By Lemma 3.2, the iterated integrals converge absolutely, so we may interchange the order of integration. We compute the $\hat{e}$ integral as follows. We write $\int_{S^2} \exp(ik\hat{e}\cdot(\vec{x}-\vec{y}))\, d\hat{e}$ in polar coordinates with the polar angle measured from the direction $\vec{x}-\vec{y}$. We can then carry out the integration explicitly: the result is

$$\int_{S^2} \exp[ik\hat{e}\cdot(\vec{x}-\vec{y})]\, d\hat{e} = -\frac{8\pi^2}{ik}[G_0^+(k,|\vec{x}-\vec{y}|) - G_0^-(k,|\vec{x}-\vec{y}|)]. \tag{3.5}$$

Finally, we use (2.3+) to obtain (3.4). □

THEOREM 3.4. *Suppose conditions* (H1±) *and* (H2) *hold for* $n=3$. *Then*

$$\psi^+(k,\hat{e}',\vec{x}) = \psi^-(k,\hat{e}',\vec{x}) + \frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')\psi^-(k,\hat{e},\vec{x})\, d\hat{e}. \tag{3.6}$$

Proof. Our plan is to obtain an equation similar to (2.3±) but with the plane wave replaced by $\exp[ik\hat{e}'\cdot\vec{x}] + ik/2\pi \int A(k,\hat{e},\hat{e}')\exp(ik\hat{e}\cdot\vec{x})\, d\hat{e}$. We will compare this new equation with (3.4); from the uniqueness of the solution of equations such as (2.3±), we will conclude (3.6).

First we multiply (2.3−) by $(ik/2\pi)A(k,\hat{e},\hat{e}')$ and integrate with respect to $\hat{e}$

$$\begin{aligned}
&\frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')\psi^-(k,\hat{e},\vec{x})\, d\hat{e} \\
&= \frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')\exp(ik\hat{e}\cdot\vec{x})\, d\hat{e} \\
&\quad + \frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')\int G_0^-(k,|\vec{x}-\vec{y}|)(V\psi^-)(k,\hat{e},\vec{y})\, d\vec{y}\, d\hat{e}.
\end{aligned} \tag{3.7}$$

We next interchange the order of integration in the last term of (3.7). This is valid because the following estimate shows that the iterated integral converges absolutely; first we use Lemma 3.2 to obtain

$$\begin{aligned}
&\int\int_{S^2} |A(k,\hat{e},\hat{e}')||G_0^-(k,|\vec{x}-\vec{y}|)||(V\psi^-)(k,\hat{e},\vec{y})|\, d\hat{e}\, d\vec{y} \\
&\leqq 4\pi \max_{S^2\times S^2} A \int |\vec{x}-\vec{y}|^{-1}|(V\psi^-)(k,\hat{e},\vec{y})|\, d\vec{y}.
\end{aligned} \tag{3.8}$$

The right side of (3.8) can easily be shown to be finite by using condition (H2) as in the proof of Lemma 3.2.

Upon interchange of the order of integrals in (3.7), we obtain

$$\frac{ik}{2\pi}\int_{S^2} A\psi^- = \frac{ik}{2\pi}\int_{S^2} A\exp + \frac{ik}{2\pi}\int G_0^- \int_{S^2} A(V\psi^-). \tag{3.9}$$

Next we write (2.3−) with $\hat{e}'$ substituted for $\hat{e}$ and add (3.9) to the result. We obtain

$$\begin{aligned}
&\psi^-(k,\hat{e}',\vec{x}) + \frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')\psi^-(k,e,\vec{x})\, de \\
&= \exp(ik\hat{e}'\cdot\vec{x}) + \frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')\exp(ik\hat{e}\cdot\vec{x})\, d\hat{e} \\
&\quad + \int G_0^-(k,|\vec{x}-\vec{y}|)\left[(V\psi^-)(k,\hat{e}',\vec{y}) - \frac{ik}{2\pi}\int_{S^2} A(k,\hat{e},\hat{e}')(V\psi^-)(k,\hat{e},\vec{y})\, d\hat{e}\right] d\vec{y}.
\end{aligned} \tag{3.10}$$

Since V is linear and is independent of $\hat{e}$, we pull V outside the bracket in the last term of (3.10). We note that by Lemma 3.2, the inhomogeneous term $\exp(ik\hat{e}'\cdot\vec{x})+ik/2\pi\int_{S^2}A(k,\hat{e},\hat{e}')\exp(ik\hat{e}\cdot\vec{x})\,d\hat{e}$ is in $H^{2,-s}$. By hypothesis (H1−), the solution of (3.10) is unique; we obtain (3.6) by comparing (3.10) to (3.4). □

4. Examples. In this section, we show that (2.6) holds for three physically relevant equations.

Example 1. *The Schrödinger equation.* Here we consider the case in which V is the operator of multiplication by a complex-valued function

$$V(\vec{x})=U(\vec{x})+iW(\vec{x}),$$

where U and W are real-valued functions satisfying hypothesis (b) of Theorem 3.1. Equation (2.1) then becomes the Schrödinger equation

$$(\nabla^2+k^2)\psi=(U+iW)\psi. \tag{4.1}$$

Equation (4.1) with nonzero W is commonly used to model many-body systems in which there is attenuation. For example it is used in the optical model for scattering from nuclei [13].

In order to show that (3.6) holds for almost all real k, we must check hypothesis (d) of Theorem 3.1.

LEMMA 4.1. *Suppose V is in $L^2(\mathbb{R}^3)$. Then for* $\operatorname{Im}k>\|V\|^2_{L^2}(32\pi^2)^{-1}$, *$I-G_0V$ is invertible on L^2.*

Sketch of proof. The square of the Hilbert-Schmidt norm [10] of G_0V can be explicitly computed for $\operatorname{Im}k>0$. It turns out to be $\|V\|^2_{L^2}(32\pi^2\operatorname{Im}k)^{-1}$. Thus for $\operatorname{Im}k>\|V\|^2_{L^2}(32\pi^2)^{-1}$, the Hilbert-Schmidt norm of G_0V is less than one, so $(I-G_0V)^{-1}$ can be constructed by iteration [10]. □

COROLLARY 4.2. *Suppose V is in L^2 and satisfies condition* (b) *of Theorem* 3.1. *Then for* $\operatorname{Im}k>\|V\|^2_{L^2}(32\pi^2)^{-1}$, $(\nabla^2+k^2-V):H^2\to L^2$ *is invertible.*

Proof. We consider the operator $G=G_0(I-G_0V)^{-1}$. It is well known [8] that for $\operatorname{Im}k>0$, G_0 maps L^2 onto H^2. This fact and Lemma 4.1 imply that G maps L^2 into H^2. By arguments similar to the one above (3.2), it can be checked that G is the inverse of (∇^2+k^2-V). □

We have proved the following result.

THEOREM 4.3. *Suppose V is in L^2 and satisfies condition* (b) *of Theorem* 3.1. *Then* (2.6) *holds for almost all real k.*

In fact, if V is real-valued, then it is known [9] *that* (H1±) *holds for all nonzero real k, and consequently* (2.6) *holds for all nonzero real k.*

Example 2. *The wave equation with variable speed.* Here we consider the equation

$$[\nabla^2+k^2c^{-2}(\vec{x})]\psi(k,\vec{x})=0, \tag{4.2}$$

where $c^{-2}(\vec{x})-1$ satisfies hypothesis (b) of Theorem 3.1. In this case, V is the operator of multiplication by $k^2(c^{-2}(\vec{x})-1)$. In order to conclude that (2.6) holds for almost all k, we must check hypothesis (d) of Theorem 3.1.

We note that it is easy to show that for $c^{-2}(\cdot)-1$ in L^2, $I-G_0V$ is invertible on L^2 for small $|k|$. (The Hilbert-Schmidt norm of G_0V can be shown to be small, and $(I-G_0V)^{-1}$ can be constructed by iteration [10].) We then define the operator $G:L^2\to H^2$ by $G=G_0(I-G_0V)^{-1}$. If $(I-G_0V)^{-1}$ exists, then G exists and is the inverse of ∇^2-V+k^2. (See the argument above (3.2).) Hypothesis (d) of Theorem 3.1 is thus satisfied, and we can conclude that (2.6) holds. We have proved the following theorem.

THEOREM 4.4. *Suppose $c^{-2}(\cdot)-1$ is in $L^2(R^3)$ and satisfies hypothesis* (b) *of Theorem* 3.1. *Then for almost all real k,* (2.6) *holds, where $V=k^2(c^{-2}(\cdot)-1)$.*

We note that c^{-2} has not been assumed to be real-valued, positive, or smooth. However, if c^{-2} satisfies the conditions of Theorem 4.4 and, in addition, is strictly positive, it is known that conditions (H1±) actually hold for all real k [9], [14], so in fact (2.6) holds for all real k.

Example 3. *The acoustic equation.* The acoustic equation governs the propagation of pressure waves in a fluid [15]

$$[\nabla^2 - \rho^{-1}\nabla\rho\cdot\nabla + k^2c^{-2}]\psi(k,\vec{x}) = 0 \tag{4.3}$$

where $\psi(k,\vec{x})$ is the pressure at $\vec{x}$, $\rho(\vec{x})$ is the density, and $c(\vec{x})$ is the speed of propagation. We will assume that for some ρ_0, $\rho(\vec{x})-\rho_0$ and $c^{-2}(\vec{x})-1$ are functions of compact support with two continuous derivatives and that $\rho(\vec{x})$ and $c^2(\vec{x})$ are positive. Thus $c^{-2}(\vec{x})-1$ and $\nabla\rho/\rho$ automatically satisfy hypotheses (b) and (c) of Theorem 3.1. Again we must check that hypothesis (d) is satisfied.

We note that (4.3) can be written

$$[-c^2\rho\nabla\cdot\rho^{-1}\nabla - k^2]\psi(k,\vec{x}) = 0.$$

It is shown in [16] that the operator $-c^2\rho\nabla\cdot\rho^{-1}\nabla$ on H^2 is unitarily equivalent to a self-adjoint operator. Therefore the spectrum of $-c^2\rho\nabla\cdot\rho^{-1}\nabla$ is contained on the real axis; this shows that $-c^2\rho\nabla\cdot\rho^{-1}\nabla - k^2$ is invertible for all complex k except possibly k on the real and imaginary axes. Thus we have shown the following theorem.

THEOREM 4.5. *Suppose that for some* ρ_0, $\rho(\vec{x})-\rho_0$ *and* $c^{-2}(\vec{x})-1$ *are functions of compact support with two continuous derivatives and that* $\rho(\vec{x})$ *and* $c^2(\vec{x})$ *are positive. Then* (2.6) *holds for almost all real* k. (*Here* $V = -\rho^{-1}\nabla\rho\cdot\nabla + k^2(c^{-2}-1)$.)

We note that when ρ is constant, (4.3) reduces to (4.2). The argument of Example 3, however, requires stronger conditions on c^2.

5. Consequences of the integral relation (2.6). Some of the consequences of (2.6) are discussed in this section. First we show that if the linear operator V satisfies (H1±) and (H2) and also has compact support (in a sense described below), then a generalized optical theorem holds for the scattering amplitude.

Next, we discuss some simple consequences of the generalized optical theorem in the long wavelength limit. In particular, for certain scatterers such as those in Example 3, we assume the scattering amplitude can be expanded in a power series about $k=0$

$$A = A_2k^2 + iA_3k^3 + \cdots.$$

The generalized optical theorem implies certain symmetry relations for the A_j.

First we use (2.6) to derive a generalized optical theorem for scattering governed by (2.1). The derivation involves using (2.4) in (2.6) and letting $|\vec{x}|\to\infty$. To be sure that (2.4) holds, we assume that the operator V has compact support

(H3) There is some radius $R>0$ such that $\chi_R V = 0$, where $\chi_R(\vec{x}) = 1$ if $|\vec{x}| > R$ and zero otherwise.

We also need the analogue of (2.4) for ψ^-:

$$\psi^-(k,\hat{e},\vec{x}) = \exp(ik\hat{e}\cdot\vec{x}) + B(k,\hat{x},\hat{e})\exp(-ik|\vec{x}|)/|\vec{x}| + O(|\vec{x}|^{-2}) \tag{5.1}$$

where

$$B(k,\hat{e},\hat{e}') = -\frac{1}{4\pi}\int \exp(-ik\hat{e}\cdot\vec{y})(V\psi^-)(k,\hat{e}',\vec{y})\,d\vec{y}. \tag{5.2}$$

THEOREM 5.1. *Suppose V satisfies* (H1±), (H2), *and* (H3) (*for* $n=3$). *Then*

$$A(k,-\hat{e},\hat{e}')-B(k,\hat{e},\hat{e}')=\frac{ik}{2\pi}\int A(k,\hat{e}'',\hat{e}')B(k,\hat{e},\hat{e}'')\,de''. \tag{5.3}$$

Proof. We substitute (2.4) in (2.6), and multiply by $|x|$. This results in

$$\begin{aligned}A(k,\hat{x},\hat{e})\exp(ik|\vec{x}|)=B(k,\hat{x},\hat{e})\exp(-ik|x|)+\frac{ik|x|}{2\pi}\int A(k,\hat{e}',\hat{e})\exp(ik\hat{e}'\cdot\vec{x})\,d\hat{e}'\\ +\frac{ik}{2\pi}\int A(k,\hat{e}',\hat{e})B(k,\hat{x},\hat{e}')\,d\hat{e}'\exp(-ik|\vec{x}|)+O(|x|^{-1}),\end{aligned} \tag{5.4}$$

where we have used Lemma 3.2. Next we consider the second term on the right side of (5.4). This term we denote by I. We write the integral in polar coordinates with $e'=(\theta,\varphi)$

$$I=\frac{ik|x|}{2\pi}\int_0^{2\pi}\int_0^{\pi}A(k,(\theta,\varphi),\hat{e})\exp(ik|\vec{x}|\cos\theta)\sin\theta\,d\theta\,d\varphi. \tag{5.5}$$

In (5.5) we integrate by parts, obtaining

$$\begin{aligned}I=A(k,\hat{x},\hat{e})\exp(ik|\vec{x}|)-A(k,-\hat{x},\hat{e})\exp(-ik|\vec{x}|)\\ -\frac{1}{2\pi}\int_0^{2\pi}\int_0^{\pi}\frac{\partial}{\partial\theta}A(k,(\theta,\varphi),\hat{e})\exp(ik|\vec{x}|\cos\theta)\,d\theta\,d\varphi.\end{aligned} \tag{5.6}$$

We next apply the method of stationary phase [17] to the θ integral of (5.6); this shows that

$$I=A(k,\hat{x},\hat{e})\exp(ik|\vec{x}|)-A(k,-\hat{x},\hat{e})\exp(-ik|\vec{x}|)+O((k|\vec{x}|)^{-1/2}). \tag{5.7}$$

We use (5.7) in (5.4), divide by $\exp(-ik|\vec{x}|)$, and let $|\vec{x}|\to\infty$. A simple relabeling of variables gives (5.3). □

We note that (5.3) simplifies if $V(-k)=V(k)$. In this case, (2.3±) together with hypotheses (H1±) show that

$$\psi^+(-k,-\hat{e},\vec{x})=\psi^-(k,\hat{e},\vec{x}). \tag{5.8}$$

This, in turn, when used with (2.4) and (5.1), shows that

$$B(k,\hat{e},\hat{e}')=A(-k,\hat{e},-\hat{e}'). \tag{5.9}$$

Equation (5.3) then becomes

$$A(k,\hat{e},\hat{e}')-A(-k,-\hat{e},-\hat{e}')=\frac{ik}{2\pi}\int A(k,\hat{e}'',\hat{e}')A(-k,-\hat{e},-\hat{e}'')\,d\hat{e}'', \tag{5.10}$$

where $-\hat{e}$ has been relabeled $\hat{e}$.

We obtain even further simplification of (5.3) if both $V(-k)=V(k)$ and $\bar{V}=V$ hold. (Here the bar denotes complex conjugate.) However, first we note that when $\bar{V}=V$, (2.3±) plus hypothesis (H1±) show that

$$\psi^-(k,\hat{e},\vec{x})=\overline{\psi^+(k,-\hat{e},\vec{x})}. \tag{5.11}$$

Equations (5.8), (5.11), and (2.4) can then be used [5] to show that

$$\overline{A(k,\hat{e},\hat{e}')}=A(-k,\hat{e},\hat{e}'). \tag{5.12}$$

Finally, (2.3±) and (2.5) can be used [5] to derive the reciprocity relation

$$A(k,\hat{e},\hat{e}')=A(k,-\hat{e}',-\hat{e}). \tag{5.13}$$

When equations (5.12) and (5.13) are used in (5.10), the result is the generalized optical theorem which is well known for some wave equations (e.g., the Schrödinger equation and equations governing electromagnetic scattering [5]),

$$A(k, \hat{e}, \hat{e}') - \overline{A(k, \hat{e}', \hat{e})} = \frac{ik}{2\pi} \int A(k, \hat{e}'', \hat{e}') \overline{A(k, \hat{e}'', \hat{e})}\, d\hat{e}''. \tag{5.14}$$

Most derivations of the optical theorem rely on unitarity, but the derivations in this paper do not. In particular, note that (5.3) differs from the usual generalized optical theorem for dissipative systems.

Next we consider wave equations and scatterers (see Examples 2 and 3) such that at long wavelength the scattering amplitude may be expanded about $k = 0$ as

$$\begin{aligned} A(k, \hat{e}, \hat{e}') = {} & A_2(\hat{e}, \hat{e}')k^2 + iA_3(\hat{e}, \hat{e}')k^3 + A_4(\hat{e}, \hat{e}')k^4 \\ & + iA_5(\hat{e}, \hat{e}')k^5 + O(k^6) \end{aligned} \tag{5.15}$$

where $A_2, A_3, \cdots$ are real functions. Substituting (5.15) into (5.14) and equating orders of k leads to the following consistency requirements:

$$A_2(\hat{e}, \hat{e}') = A_2(-\hat{e}, -\hat{e}'), \tag{5.16}$$

$$A_3(\hat{e}, \hat{e}') = -A_3(-\hat{e}, -\hat{e}'), \tag{5.17}$$

$$A_4(\hat{e}, \hat{e}') = A_4(-\hat{e}, -\hat{e}'), \tag{5.18}$$

$$A_5(\hat{e}, \hat{e}') + A_5(-\hat{e}, -\hat{e}') = \frac{1}{2\pi} \int d^2\hat{e}'' A_2(\hat{e}'', \hat{e}) A_2(\hat{e}'', \hat{e}'). \tag{5.19}$$

Note that these relations do *not* follow solely from reciprocity. For a number of circumstances (5.17) implies that $A_3(\hat{e}, \hat{e}') = 0$. This is important in practical problems because it allows signals from different transducers to be matched together properly. This will be discussed in more detail in the next section.

6. Applications to inverse scattering. In this section we will discuss the applications of (2.6) and (5.15)–(5.17) to inverse scattering. First we will show that (5.17) can be used in an important phase retrieval problem. Then we will show how (2.6) can be used to find V from A. Finally, we conclude with a few remarks on other applications to inverse scattering.

The phase retrieval problem arises as follows. In applications of inverse scattering theory to experiments (e.g., various forms of imaging), the scattering amplitude is only known in a corrupted form. Generally it is necessary to solve various preliminary problems in order to obtain the scattering amplitude from measured data. Due to measurement error and difficulties in the precise mathematical modeling of the physical process, it often happens that the scattering amplitude can be found only up to a phase factor which depends on the directions of incidence and scattering. Proceeding without correcting for this unknown phase factor leads to "blurring" in the image produced by the inverse scattering algorithm. See [19] for a discussion of the physical difficulties.

Consequently the following problem, which is an idealization of the physical system just discussed, is of interest. Suppose we know the corrupted form A^c of the scattering amplitude A, where

$$A^c(k, \hat{e}, \hat{e}') = A(k, \hat{e}, \hat{e}') \exp[ik\tau(\hat{e}, \hat{e}')], \tag{6.1}$$

τ being a bounded, real function that is otherwise unknown. Then the problem is to recover A given A^c. In general this problem is impossible to solve, since both A and τ are unknown.

However, let us further suppose that the governing wave equation satisfies (5.15)–(5.17). In addition, let us suppose that the scatterer is known to be symmetric with respect to the origin. We show below that $A_3(\hat{e}, \hat{e}')$ vanishes, $A_2(\hat{e}, \hat{e}') = A_2^c(\hat{e}, \hat{e}')$ and

$$A(k, \hat{e}, \hat{e}') = A^c(k, \hat{e}, \hat{e}') \exp [-ikA_3^c(\hat{e}, \hat{e}')/A_2^c(\hat{e}, \hat{e}')]. \tag{6.2}$$

Here we have expanded the right-hand side of (6.1) as a power series in k, which implies that A^c can be expanded as

$$A^c(k, \hat{e}, \hat{e}') = k^2 A_2^c(\hat{e}, \hat{e}') + ik^3 A_3^c(\hat{e}, \hat{e}') + \mathcal{O}(k^4), \tag{6.3}$$

where the expansion coefficients are real.

We show (6.2) by noting that the inversion symmetry implies that $A(k, \hat{e}, \hat{e}') = A(k, -\hat{e}, -\hat{e}')$ and consequently that

$$A_3(\hat{e}, \hat{e}') = A_3(-\hat{e}, -\hat{e}'). \tag{6.4}$$

Comparison of (6.4) and (5.17) shows that $A_3 = 0$. Now expand both sides of (6.1) in a power series in k and equate coefficients of corresponding powers of k. We obtain $A_2 = A_2^c$ and $\tau = A_3^c/A_2^c$. Substitution of these results in (6.1) and rearrangement yield (6.2).

Equation (6.2) solves the phase retrieval problem: it shows how the true scattering amplitude can be recovered from the corrupted one.

Equation (2.6) can also be used to attack the inverse scattering problem. This problem is to determine V from $A(k, e, e')$. One difficulty in using (2.6) for this is that (2.6) contains two unknowns ψ^+ and ψ^-. In order to obtain an equation with only one unknown, it is generally necessary to Fourier transform and use domain of dependence results to separate ψ^+ and ψ^-. This procedure results in an integral equation for the wavefield in terms of the scattering data.

Specifically, the theory goes as follows in the Schrödinger equation case. We denote the Fourier transform of $\psi^\pm$ by

$$u^\pm(t, \hat{e}, \vec{x}) = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(-ikt)\psi^\pm(k, \hat{e}, \vec{x})\, dk \tag{6.5}$$

and the Fourier transform of $A(k, \hat{e}, \hat{e}')$ by

$$R(t, \hat{e}, \hat{e}') = (2\pi)^{-1} \int_{-\infty}^{\infty} \exp(-ikt)A(k, \hat{e}, \hat{e}')\, dk. \tag{6.6}$$

In this notation, the Fourier transform of (2.6) is

$$u^+(t, \hat{e}, \vec{x}) = u^-(t, \hat{e}, \vec{x}) - (2\pi)^{-1} \int_{-\infty}^{\infty} \int_{S^2} \dot{R}(t-\tau, \hat{e}', \hat{e}) u^-(\tau, \hat{e}', \vec{x})\, d\hat{e}'\, d\tau, \tag{6.7}$$

where the dot denotes differentiation with respect to t. We note that u^+ satisfies the hyperbolic equation

$$(\nabla^2 - \partial_{tt} - V)u^+ = 0 \tag{6.8}$$

together with the condition that for large negative times, $u^+(t, \hat{e}, \vec{x}) = \delta(t - \hat{e} \cdot \vec{x})$. By domain-of-dependence results for (6.8), when V is real, u^+ satisfies

$$u^+(t, \hat{e}, \vec{x}) = 0 \quad \text{for } t < \hat{e} \cdot \vec{x}. \tag{6.9}$$

In addition, by (5.8),

$$u^-(t, \hat{e}, \vec{x}) = u^+(-t, -\hat{e}, \vec{x}) \tag{6.10}$$

and therefore

$$u^-(t, \hat{e}, \vec{x}) = 0 \quad \text{for } t > \hat{e} \cdot \vec{x}. \tag{6.11}$$

Writing $u^{\pm} = \delta + u^{\pm}_{sc}$ and taking (6.10) and (6.11) into account, (6.7) for $t > \hat{e} \cdot \vec{x}$ becomes

$$\begin{aligned} u^+_{sc}(t, \hat{e}, \vec{x}) = &-(2\pi)^{-1} \int_{S^2} \dot{R}(t - \hat{e}' \cdot \vec{x}, \hat{e}', \hat{e})\, d\hat{e}' \\ &-(2\pi)^{-1} \int_{-\infty}^{\infty} \int_{S^2} \dot{R}(t + \tau, -\hat{e}', \hat{e}) u^+_{sc}(\tau, \hat{e}', \vec{x})\, d\hat{e}'\, d\tau. \end{aligned} \tag{6.12}$$

Newton has shown [1], [20] that (6.12) is a Fredholm equation that can always be solved. We thus obtain u^+_{sc}, and from it, V can be obtained via the formula [1], [21]

$$V(\vec{x}) = -2\hat{e} \cdot \nabla u^+_{sc}(\hat{e} \cdot \vec{x}, \hat{e}, \vec{x}). \tag{6.13}$$

For more details concerning this method of solving the inverse scattering problem for the Schrödinger equation, the reader is referred to [1], [6], and [20].

It is natural to try a similar procedure for solving the inverse scattering problem for the wave equation (Example 2). For this equation, the inverse problem is to recover $c(\vec{x})$ from $A(k, \hat{e}, \hat{e}')$. Again we start with (2.6) which is Fourier transformed [14] to obtain (6.7). The next step is to use causality; for the wave equation, (6.9) holds only when $c(\vec{x}) \leqq 1$ for all $\vec{x}$. Thus we obtain (6.12) only under the hypothesis that $c(\vec{x}) \leqq 1$.

Equation (6.12) is no longer a Fredholm equation, and little is known about it. Nevertheless, if it can be solved for u^+, then $c(\vec{x})$ can be recovered as follows. We know from geometrical optics that

$$u^+(t, \hat{e}, \vec{x}) = z(\hat{e}, \vec{x}) \delta(t - s(\hat{e}, \vec{x})) + (\text{less singular terms}). \tag{6.14}$$

If u^+ is known, then $s(\hat{e}, \vec{x})$ is known; $c(\vec{x})$ is related to s by

$$c^{-2}(\vec{x}) = |\nabla s(\hat{e}, \vec{x})|^2. \tag{6.15}$$

For more details concerning this method of solving the inverse scattering problem for the wave equation, the reader is referred to [2].

The wave equation inverse scattering problem can also be attacked with (2.6) directly. Equations (4.2) and (2.6), together with

$$V(\vec{x}) = \nabla^2[(\psi^+(k, \hat{e}, \vec{x}) - \exp(ik\hat{e} \cdot \vec{x}))/k^2]_{k=0}, \tag{6.16}$$

form a system of equations whose simultaneous solution solves the inverse scattering problem. This system of equations can be attacked by an iterative method; preliminary numerical results indicate that this method converges for some problems. More details about this can be found in [23] and [14].

Finally, relation (2.6) for the wave equation can also be used together with various approximations. For example, if $k^2 V$ is small, we expect to recover from (2.6) the Born approximation and inverse methods based on it. Similarly, if k is large but V is small, then we expect that geometrical optics should be useful and that the rays should be nearly straight. Some such approximate schemes have been worked out in [22].

REFERENCES

[1] R. G. NEWTON, *Inverse scattering.* II. *Three dimensions*, J. Math. Phys., 21 (1980), pp. 1698-1715. See also R. G. Newton, *The Marchenko and Gelfand-Levitan methods in the inverse scattering problem in one and three dimensions*, in Conference on Inverse Scattering: Theory and Application, J. B. Bednar, R. Redner, E. Robinson, and A. Weglein, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983, and references therein.

[2] J. H. ROSE, M. CHENEY, AND B. DEFACIO, *Three-dimensional inverse scattering: plasma and variable velocity wave equations*, J. Math. Phys., 26 (1985), pp. 2803–2813.
[3] P. D. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.
[4] E. G. SCHMIDT, *On the representation of the potential scattering operator in quantum mechanics*, J. Differential Equations, 7 (1970), pp. 389–394.
[5] R. G. NEWTON, *Scattering Theory of Waves and Particles*, 2nd ed., Springer-Verlag, New York, 1982.
[6] J. H. ROSE, M. CHENEY, AND B. DEFACIO, *The connection between time- and frequency-domain three-dimensional inverse scattering methods*, J. Math. Phys., 25 (1984), pp. 2995–3000.
[7] ———, *Determination of the wavefield from the scattering data*, Phys. Rev. Lett., 57 (1986), pp. 783–786.
[8] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics. II. Fourier Analysis and Self-Adjointness*, Academic Press, New York, 1975.
[9] S. AGMON, *Spectral properties of Schrödinger operators and scattering theory*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 2 (1975), pp. 151–218.
[10] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics. I: Functional Analysis*, Academic Press, New York, 1972.
[11] B. SIMON, *Quantum Mechanics for Hamiltonians Defined as Quadratic Forms*, Princeton University Press, Princeton, NJ, 1971.
[12] M. SCHECHTER, *Spectra of Partial Differential Operators*, North-Holland, New York, 1971.
[13] P. H. HODGSON, *Nuclear Reactions and Nuclear Structure*, Clarendon Press, Oxford, 1972.
[14] M. CHENEY AND E. SOMERSALO, *Estimates for wave propagation in inhomogeneous media*, unpublished manuscript.
[15] P. M. MORSE AND K. U. INGARD, *Theoretical Acoustics*, McGraw-Hill, New York, 1968.
[16] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics. III: Scattering Theory*, Academic Press, New York, 1979.
[17] E. T. COPSON, *Asymptotic Expansions*, Cambridge University Press, New York, 1965.
[18] J. M. RICHARDSON, *Scattering of elastic waves from symmetric inhomogeneities at low frequencies*, Wave Motion, 6 (1984), pp. 325–336.
[19] J. H. ROSE, *Phase retrieval for the variable velocity classical wave equation*, Inverse Problems, 2 (1986), pp. 219–228.
[20] R. G. NEWTON, *Variational principles for inverse scattering*, Inverse Problems, 1 (1985), pp. 371–380.
[21] M. CHENEY, *A rigorous derivation of the "miracle" identity of three-dimensional inverse scattering*, J. Math. Phys., 25 (1984), pp. 2988–2990.
[22] M. CHENEY AND J. H. ROSE, *Three-dimensional inverse scattering for the wave equation: low frequency approximations with error estimates*, Inverse Problems, in press.
[23] J. H. ROSE AND M. CHENEY, *Self-consistent equations for variable velocity three-dimensional inverse scattering*, Phys. Rev. Lett., 59 (1987), pp. 954–957.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

006

A SINGULAR LIMIT PROBLEM FOR A VOLTERRA EQUATION*

RICHARD NOREN†

Abstract. Concerning the solution $u(t, c)$ of the equation

$$u'(t)+\int_0^t [a(t-s)+c]u(s)\,ds=0, \qquad u(0)=1$$

weaker sufficient conditions are found for $\int_0^\infty \sup_{0\leqq c\leqq 1}|u(t, c)|\,dt<\infty$ than were previously known. In particular the assumption $(-1)^k a^{(k)}(t)\geqq 0$, $t>0$, $k=0, 1, 2, \cdots$, is replaced by the assumption a is nonnegative, nonincreasing, convex and $-a'$ is convex for $t>0$.

Key words. Volterra equation, L^1, completely monotone, nonnegative, nonincreasing, convex

AMS(MOS) subject classification. 45

1. Introduction.

Concerning the solution $u=u(t)=u(t, c)$ of the equation

$$u'(t)+\int_0^t [a(t-s)+c]u(s)\,ds=0, \quad t\geqq 0, \quad u(0)=1, \quad \left(' =\frac{d}{dt}\right), \tag{1.1}$$

we prove the following theorem.

THEOREM 1. *Let $a(t)$ satisfy*

$$a\in L^1(0, 1) \text{ is nonnegative, nonincreasing, convex and } -a' \text{ is convex on } (0, \infty);\ 0=a(\infty)<a(0+)\leqq\infty, \tag{1.2}$$

and

$$\int_1^\infty \frac{\log u}{uA(u)}\,du<\infty, \tag{1.3}$$

where $A(u)\equiv\int_0^u a(s)\,ds$. Then

$$\int_0^\infty \sup_{0\leqq c\leqq 1}|u(t, c)|\,dt<\infty. \tag{1.4}$$

Define $\hat{a}(\tau)\equiv\int_0^\infty e^{-i\tau t}a(t)\,dt\equiv\phi(\tau)-i\tau\theta(\tau)$, where the integral exists for $\operatorname{Im}\tau<0$, *$\hat{a}$ is extended by continuity to* $(\operatorname{Im}\tau\leqq 0,\ \tau\neq 0)$ *and $\hat{a}\in C^1(0, \infty)$.*

In [7], (1.4) is proved for the completely monotonic function $a(t)=\int_0^\infty e^{-xt}\,d\alpha(x)$ where it is assumed that the nondecreasing function $\alpha(t)$ $(0\leqq t<\infty)$ also satisfies

$$\int_0^{x_0}\frac{dx}{x\alpha'(x)}<\infty \quad \text{for some } x_0>0. \tag{1.5}$$

The proof depends on the decomposition of u into a completely monotonic term plus an exponentially decaying term as given in [6] and on obtaining certain estimates uniformly in $0\leqq c\leqq 1$.

* Received by the editors September 22, 1986; accepted for publication August 10, 1987.
† Department of Mathematical Sciences, Old Dominion University, Norfolk, Virginia 23508.

In our proof of Theorem 1 we use techniques similar to those used in [1] and the following inequalities. Assuming (1.2),

$$2^{-3/2}A(\tau^{-1}) \leqq |\hat{a}(\tau)| \leqq 4A(\tau^{-1}), \qquad \tau > 0, \tag{1.6}$$

$$\frac{1}{5}A_1(\tau^{-1}) \leqq \theta(\tau) \leqq 12A_1(\tau^{-1}), \qquad \tau > 0 \qquad \left(A_1(x) \equiv \int_0^x sa(s)\,ds\right), \tag{1.7}$$

$$|\hat{a}'(\tau)| \leqq 40A_1(\tau^{-1}), \qquad |\hat{a}''(\tau)| \leqq 6000\int_0^{1/\tau} r^2 a(r)\,dr, \qquad \tau > 0 \tag{1.8}$$

(see [1, (4.1), (4.3), (4.2), and (5.3)]),

$$|u(t, c)| \leqq 1, \qquad 0 \leqq t < \infty, \qquad 0 \leqq c \tag{1.9}$$

(see [3, Thm. 2] and [8]. The number $\sqrt{2}$ appears in [3] instead of 1 because of a typographical error). Note that (1.6)-(1.9) all hold without the assumption $-a'$ is convex except for the second inequality in (1.8).

In [1], [4], [5] the inequality analogous to (1.4),

$$\int_0^\infty \sup_{1 \leqq \lambda < \infty} |u_\lambda(t)|\,dt < \infty,$$

is obtained under the assumption (1.2) for the problem

$$u_\lambda'(t) + \lambda \int_0^t [a(t-s)+d]u_\lambda(s)\,ds = 0, \qquad u_\lambda(0) = 1,$$

with d fixed and nonnegative.

For (1.4) to hold it is necessary that $a(t) \notin L^1(0, \infty)$ (see [7, p. 200]) but not sufficient as the completely monotonic function $a(t) = (1-e^{-t})/t$ shows (see [7, § 4]). Clearly, (1.3) implies that $a \notin L^1(0, \infty)$.

The conditions (1.3) and (1.5) are similar. Thus if $\alpha'(x) = (-\log x)^q$ for $0 < x < x_0$, $q > 0$, then $a(t)$ behaves like $t^{-1}\log^q t$ as $t \to \infty$, as the following two calculations show. First, we have

$$t^{-1}\log^q t \int_0^1 e^{-x}\,dx < t^{-1}\int_0^1 e^{-x}\log^q(t/x)\,dx \leqq t^{-1}\int_0^{tx_0} e^{-x}\log^q(t/x)\,dx$$

$$= \int_0^{x_0} e^{-yt}(-\log y)^q\,dy \leqq a(t), \qquad x_0^{-1} < t.$$

Also for $0 < \varepsilon < x_0$, we have

$$\begin{aligned}
a(t) &= \int_0^{x_0} e^{-yt}(-\log y)^q\,dy + \int_{x_0}^\infty e^{-yt}\alpha'(y)\,dy \\
&\leqq t^{-1}\int_0^{tx_0} e^{-x}\log^q(t/x)\,dx + e^{-\varepsilon t}\int_{x_0}^\infty e^{-t(y-\varepsilon)}\alpha'(y)\,dy \\
&\leqq M_1\left[t^{-1}\int_0^{tx_0} e^{-x}(|\log^q t| + |\log^q x|)\,dx + e^{-\varepsilon t}\int_{x_0-\varepsilon}^\infty e^{-tx}\alpha'(x+\varepsilon)\,dx\right] \\
&\leqq M_2\left[t^{-1}\log^q t\int_0^\infty e^{-x}(1+|\log^q x|)\,dx + e^{-\varepsilon t}t\int_{(x_0-\varepsilon)t}^\infty e^{-u}\alpha(u/t+\varepsilon)\,du\right] \\
&\leqq M_3\left[t^{-1}\log^q t + e^{-\varepsilon t}t\int_{x_0-\varepsilon}^\infty e^{-u}\alpha(u+\varepsilon)\,du\right] \\
&\leqq M_4 t^{-1}\log^q t \quad \text{for } t > \max\left\{\frac{1}{x_0}, 1\right\} \qquad (M_1, M_2, M_3, M_4, \text{constant}),
\end{aligned}$$

where the first inequality uses $x = yt$, the second inequality uses $(x+y)^q \leqq 2^q(x^q+y^q)$ for $x, y, q > 0$ and $y = x+\varepsilon$, the third inequality uses integration by parts, the fourth inequality uses the fact α is positive and nondecreasing and $t > 1$, and the last inequality uses the fact that $a \in L^1(0,1)$ if and only if $\int_{x_0}^{\infty} \alpha(x)x^{-2}\,dx < \infty$, and therefore

$$\int_{x_0-\varepsilon}^{\infty} e^{-u}\alpha(u+\varepsilon)\,du = e^{\varepsilon}\int_{x_0}^{\infty} e^{-x}\alpha(x)\,dx < \infty.$$

In this case (1.3) and (1.5) both hold if and only if $q > 1$. Also, $a(t) = \Gamma(p)t^{-p}$ corresponds to $\alpha'(x) = x^{p-1}$, $p > 0$ and (1.3) and (1.5) both hold if and only if $0 < p < 1$. For completely monotonic a, condition (1.5) rules out purely jump functions α, whereas (1.3) does not.

2. Proof of Theorem 1. By [9, Thm. 2] we have $\int_0^{\infty} |u(t,c)|\,dt < \infty$ for $c \geqq 0$. We will prove that

$$\int_0^{\infty} \sup_{0 \leqq c \leqq 1} |u(t,c) - u(t,0)|\,dt < \infty,$$

which implies (1.4) by the triangle inequality. Let $D \equiv D(\tau) \equiv \hat{a}(\tau) + i\tau$, $D(\tau, c) \equiv D - ic\tau^{-1}$. By [1, (4.32)], we have the representation

$$\begin{aligned}
\pi u(t,c) - \pi u(t,0) &= \operatorname{Im}\left\{\frac{1}{t}\int_0^{\infty} e^{i\tau t}\left[\frac{D_\tau(\tau,c)}{D(\tau,c)^2} - \frac{D'}{D^2}\right]d\tau\right\} \\
&= \operatorname{Im}\left\{\frac{1}{t}\int_0^{\infty} e^{i\tau t}\left[\frac{ic\tau^{-2}}{D(\tau,c)^2} + \frac{2ic\tau^{-1}D'}{DD(\tau,c)^2} + \frac{c^2\tau^{-2}D'}{D^2D(\tau,c)^2}\right]d\tau\right\} \qquad (2.1) \\
&\equiv \operatorname{Im}\left\{\frac{1}{t}\left(\int_0^{1/t} + \int_{1/t}^{\varepsilon} + \int_{\varepsilon}^{K} + \int_K^{\infty}\right) e^{i\tau t}[I_1 + I_2 + I_3]\,d\tau\right\}, \qquad t > 0,
\end{aligned}$$

where ε and K are constants that will be defined in the next paragraph.

To estimate the right-hand side of (2.1) we need lower bounds for $|D(\tau,c)|$ (and for $|D| = |D(t,0)|$). We first note that (1.7) and $a \notin L^1(1,\infty)$ imply that, for some $\varepsilon > 0$,

$$2\tau \leqq \tau\theta(\tau), \qquad 0 < \tau \leqq \varepsilon, \qquad (2.2)$$

and then by (1.6) (for $0 < \tau \leqq \varepsilon$, $0 \leqq c \leqq 1$)

$$\begin{aligned}
|D(\tau,c)|^2 &= \phi^2(\tau) + (c\tau^{-1} + \tau\theta(\tau) - \tau)^2 \geqq \phi^2(\tau) + (c\tau^{-1} + \tau\theta(\tau)/2)^2 \\
&\geqq \max\{c^2\tau^{-2}, |\hat{a}(\tau)|^2/4\} \geqq \max\{c^2\tau^{-2}, A^2(\tau^{-1})/32\}.
\end{aligned}$$

That is

$$|D(\tau,c)| \geqq \max\{c\tau^{-1}, A(\tau^{-1})/\sqrt{32}\}, \qquad 0 < \tau \leqq \varepsilon, \quad 0 \leqq c \leqq 1. \qquad (2.3)$$

Also by (1.6) we see that there exists a constant $K > \max\{1, \varepsilon\}$ so that

$$|D(\tau,c)| \geqq \tau - 1, \qquad \tau \geqq K. \qquad (2.4)$$

We will use these estimates with (2.1) to prove that

$$\sup_{0 \leqq c \leqq 1} |u(t,c) - u(t,0)| \leqq f(t), \qquad t > 1/\varepsilon,$$

where $\int_{1/\varepsilon}^{\infty} f(t)\,dt < \infty$. This together with (1.9) will complete the proof. The function f that we use is defined by

$$f(t) \equiv M\left[\frac{1}{t^2} + \frac{1}{tA(t)} + \frac{1}{t}\int_t^{\infty} \frac{du}{uA(u)} + \frac{1}{t^2}\int_{1/\varepsilon}^{t} \frac{du}{A(u)}\right].$$

(Here and for the rest of the paper M denotes a constant whose exact value may change each time it appears.) A simple calculation that uses the Fubini Theorem and (1.3) shows that $\int_{1/\varepsilon}^{\infty} f(t)\,dt < \infty$.

Let $t > 1/\varepsilon$. By (2.3) and (1.8) we have

$$\left|\operatorname{Im}\left\{\frac{1}{t}\int_0^{1/t} e^{i\tau t}[I_1+I_2+I_3]\,d\tau\right\}\right| \leqq \frac{M}{t}\int_0^{1/t}\frac{\tau^{-1}}{A(\tau^{-1})}+\frac{A_1(\tau^{-1})}{A^2(\tau^{-1})}\,d\tau$$
$$\leqq \frac{M}{t}\int_0^{1/t}\frac{\tau^{-1}}{A(\tau^{-1})}\,d\tau$$
$$= \frac{M}{t}\int_t^{\infty}\frac{du}{uA(u)} \leqq f(t).$$

Now we integrate by parts and use (2.3) and (1.8) to obtain

$$\left|\operatorname{Im}\left\{\frac{1}{t}\left\{\int_{1/t}^{\varepsilon} e^{i\tau t}I_1\,d\tau\right\}\right\}\right| = \left|\frac{1}{t^2}\operatorname{Im}\left\{\frac{c\,e^{i\varepsilon t}\varepsilon^{-2}}{D(\varepsilon,c)^2}-\frac{e^{i}t^2c}{D(1/t,c)^2}\right.\right.$$
$$\left.\left.+2\int_{1/t}^{\varepsilon} e^{i\tau t}\left(\frac{\tau^{-3}c}{D(\tau,c)^2}+\frac{c\tau^{-2}D_\tau(\tau,c)}{D(\tau,c)^3}\right)d\tau\right\}\right|$$
$$\leqq \frac{M}{t^2}\left[1+\frac{t}{A(t)}+\int_{1/t}^{\varepsilon}\frac{\tau^{-2}}{A(\tau^{-1})}\,d\tau\right]$$
$$= M\left[\frac{1}{t^2}+\frac{1}{tA(t)}+\frac{1}{t^2}\int_{1/\varepsilon}^{t}\frac{du}{A(u)}\right] \leqq f(t).$$

The terms $-\operatorname{Im}\{1/t\int_{1/t}^{\varepsilon} e^{i\tau t}(I_2+I_3)\,d\tau\}$ are treated in exactly the same way.

Again we integrate by parts, then use $|D(\tau,c)| \geqq \phi(\tau) \geqq m > 0$ (m constant), $\varepsilon \leqq \tau \leqq K$, and (1.8) to obtain

$$\left|\operatorname{Im}\left\{\frac{1}{t}\int_{\varepsilon}^{K} e^{i\tau t}I_1\,d\tau\right\}\right| = \left|\frac{1}{t^2}\operatorname{Im}\left\{\frac{c\,e^{iKt}K^{-2}}{D(K,c)^2}-\frac{c\,e^{i\varepsilon t}\varepsilon^{-2}}{D(\varepsilon,c)^2}\right.\right.$$
$$\left.\left.+2\int_{\varepsilon}^{K} e^{i\tau t}\left(\frac{\tau^{-3}c}{D(\tau,c)^2}+\frac{c\tau^{-2}D_\tau(\tau,c)}{D(\tau,c)^3}\right)d\tau\right\}\right|$$
$$\leqq \frac{M}{t^2} \leqq f(t).$$

The terms $-\operatorname{Im}\{1/t\int_{\varepsilon}^{K} e^{i\tau t}(I_2+I_3)\,d\tau\}$ are also treated in this way.

Finally we integrate by parts and use (2.4) and (1.8) to obtain

$$\left|\frac{1}{t}\operatorname{Im}\left\{\int_K^{\infty} e^{i\tau t}I_1\,d\tau\right\}\right| \leqq \frac{M}{t^2}+\frac{M}{t^2}\int_K^{\infty}\frac{\tau^{-3}}{(\tau-1)^2}+\frac{\tau^{-2}}{(\tau-1)^3}\,d\tau$$
$$\leqq \frac{M}{t^2} \leqq f(t).$$

The terms $\operatorname{Im}\{1/t\int_K^{\infty} e^{i\tau t}(I_2+I_3)\,d\tau\}$ are treated in this way also. This completes the proof.

REFERENCES

[1] R. W. Carr and K. B. Hannsgen, *A nonhomogeneous integrodifferential equation in Hilbert space*, SIAM J. Math. Anal., 10 (1979), pp. 961-984.

[2] K. B. HANNSGEN, *Indirect Abelian theorems and a linear Volterra equation*, Trans. Amer. Math. Soc., 142 (1969), pp. 539-555.

[3] ———, *A Volterra equation with parameter*, SIAM J. Math. Anal., 4 (1973), pp. 22-30.

[4] ———, *The resolvent kernel of an integrodifferential equation in Hilbert space*, SIAM J. Math. Anal., 7 (1976), pp. 481-490.

[5] ———, *Uniform L^1 behavior for an integrodifferential equation with parameter*, SIAM J. Math. Anal., 8 (1977), pp. 626-639.

[6] K. B. HANNSGEN AND R. L. WHEELER, *Complete monotonicity and resolvents of Volterra integrodifferential equations*, SIAM J. Math. Anal., 13 (1982), pp. 962-969.

[7] ———, *A singular limit problem for an integrodifferential equation*, J. Integral Equations, 5 (1983), pp. 199-209.

[8] J. J. LEVIN, *The asymptotic behavior of the solution of a Volterra equation*, Proc. Amer. Math. Soc., 14 (1963), pp. 534-541.

[9] D. F. SHEA AND S. WAINGER, *Variants of the Wiener-Levy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312-343.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

007

A LOTKA–McKENDRICK MODEL FOR A POPULATION STRUCTURED BY THE LEVEL OF PARASITIC INFECTION*

R. WALDSTÄTTER†, K. P. HADELER†, AND G. GREINER‡

Abstract. A population is subdivided into classes of noninfected and infected individuals. The latter is structured by a real variable measuring the level of infection. Birth, death, and immigration of parasites are modeled by a diffusion operator, birth of hosts by a Lotka–McKendrick birth law. The resulting evolution equation is treated by semigroup methods. The generator is interpreted as a self-adjoint operator with a one-dimensional unbounded perturbation.

Key words. epidemics, one parameter semigroups, self-adjoint operator

AMS(MOS) subject classifications. 92A15, 47A55, 47B25

1. Introduction. The classical epidemic model of Kermack–McKendrick and its extensions (SIR, SIS, etc.), formulated in terms of ordinary differential equations, describes the development of an epidemic disease under three important assumptions. (1) The individuals are classified as infected or noninfected; there is no subdivision of the infected population according to degree of illnesss, degree of infectivity, number of parasites. (2) The disease is transmitted by direct contact between individuals. (3) The parasite population is acquired at one instant.

For many infectious diseases these assumptions are quite appropriate. Many diseases caused by bacteria and viruses that are or at least have been widespread in human populations are transmitted by direct contact; the parasite population is acquired at one instant and is large and unstructured. On the other hand, in diseases caused by macroparasites, in particular helminthic diseases such as onchocerciasis, the parasite population within one host contains few individuals that are acquired at different times and die at different times. Typically in the life cycle of these parasites, there are states in intermediate hosts; these hosts may act as vectors for the larvae of the parasites or the larvae are acquired from the environment.

Anderson and May [1], [2] have designed and investigated models for such diseases with the assumption of a finite number of parasites per host. In [1], [2], and successive papers, a priori assumptions have been made about the distribution of parasites within hosts; moments of such distributions have been introduced into ordinary differential equations which then could be investigated by phase plane methods. In [3] and in subsequent papers, it is assumed that the parasite population within a host is governed by a birth and death process with the killing of the host. With a linearity assumption, this model leads to first-order partial differential equations for generating functions.

If the number of parasites per host is large but fluctuating according to birth, death, or immigration of individual parasites, it appears justified to introduce a continuous variable $x \geqq 0$ measuring the size of the parasite population within a host and to model the development of the host and parasite populations by a diffusion equation.

* Received by the editors April 3, 1987; accepted for publication August 13, 1987.

† Lehrstuhl für Biomathematik, Universität Tübingen, Auf der Morgenstelle 10, D-7400 Tübingen, Federal Republic of Germany.

‡ Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-7400 Tübingen, Federal Republic of Germany.

The typical diffusion approximations for birth and death processes on a half-line $x \geqq 0$ yield differential operators with a singularity at $x = 0$. Then the solutions of the evolution equation tend to infinity for $x \to 0$. In epidemiological terms: all individuals are infected, though many to a very low extent. Such models have at least two disadvantages. (i) There is no well-defined class of "susceptibles" or "noninfected." Such a class can be defined only by arbitrarily prescribing a level of infection below which the infection is considered negligible or tolerable. (ii) The offspring of individuals of any parasite load should be noninfected (so-called vertical transmission is excluded). This requirement is difficult to realize if there is no mass at $x = 0$.

For these reasons in the announcement [5] the following approach has been chosen. There is a class of definitely noninfected individuals and a class of infected individuals that are classified according to the level of infection $x \geqq 0$. In other words, we have introduced a distribution on the half-line $x \geqq 0$ together with a point mass at zero. Furthermore, we assume that the distribution of infected individuals is continuous on $x \geqq 0$.

We assume that, apart from stochastic birth and death of parasites, the hosts acquire parasites according to a conservation law

$$u_t = (\tilde{\phi}(x)u)_x$$

where $\tilde{\phi}$ is a function describing the presence of parasites in the environment and the reaction of the host.

It turns out that with these assumptions the structure of the model is essentially determined.

If we want to interpret the model as a diffusion approximation, the assumptions of a class of noninfected and of a distribution of infected continuous at $x = 0$ may lead to technical difficulties. It seems unavoidable in modeling a population structured by a continuous level x of infection that difficulties of interpretation arise near $x = 0$.

The nucleus of the model is a diffusion equation on a half-line associated with an ordinary differential equation. This situation is similar to diffusion in a long narrow tube connected to a reservoir. Such diffusion and heat conduction problems lead to self-adjoint operator equations in suitable spaces. Hence it is not surprising that methods from the theory of self-adjoint operators are useful for the present model.

A complete discussion of the model should comprise the proof of existence and uniqueness of the initial value problem, the characterization of stationary ("persistent") solutions and their stability, and, furthermore the quantitative behavior of the solutions for some realistic choices of the parameters.

Compared to these goals the results of the present paper are somewhat modest. We show that the linear part of the model, for a constant infection rate, has a solution in some suitable Hilbert space. The mathematical approach to this linear problem seems sufficiently novel and interesting to present here.

In [10] it has been proved that the evolution equations preserve positivity, i.e., nonnegative initial data give rise to nonnegative solutions for $t \geqq 0$.

2. Description of the model. In a first step we describe the variables and parameters of the model. Let t be the chronological time and x the individual parasite load or the level of infection of a host. Let $u(t, x)$ be the density of infected hosts at time t, i.e.,

$$\int_{x_1}^{x_2} u(t, x)\, dx$$

is the number of infected hosts at time t with a level of infection between x_1 and x_2.

Let $U(t)$ be the number of noninfected hosts at time t. Then

$$P(t) = U(t) + \int_0^\infty u(t, x)\, dx \tag{1}$$

is the total population size of hosts at time t. The model assumes a distinction between noninfected hosts and hosts with a very low ($x = 0$) level of infection. It would not be justified to require $U(t) = u(t, 0)$, since U is the size of a compartment and $u(t, 0)$ is one value of a density.

Let the functions $k(x)$ and $l(x)$ describe the stochastic birth and death within the parasite population, $k(x)$ is the diffusion rate, and $l(x)$ is the drift coefficient of the diffusion process governing the change of the level of infection within hosts. The function $k(x)$ is positive; the function $l(x)$ can assume either sign.

The function $\tilde{\varphi}(t, x)$ describes the rate at which new parasites are acquired by hosts with infection level x at time t. Of course $\tilde{\varphi}(t, x) \geqq 0$. We shall assume that $\tilde{\varphi}$ can be represented as $\tilde{\varphi}(t, x) = \varphi(t)\phi(x)$, where the function $\phi(x)$ describes the reaction of the host and the function $\varphi(t)$ depends on the presence of infected vectors. The latter quantity will in turn depend on the average parasite load of the population. Similarly the function $\tilde{\varphi}_0(t)$ is the rate at which noninfected hosts acquire parasites. For reasons of symmetry we write $\tilde{\varphi}_0(t) = \varphi(t) \cdot \phi_0$ with the same $\varphi(t)$ as before. Here $\phi_0 \geqq 0$ is a constant.

The parameter γ enters the boundary condition of the diffusion equation at $x = 0$. It is the rate at which infected hosts at low infection rates lose all their parasites and become noninfected.

So far the parameters mainly describe the changes of the parasite population within the hosts. The dynamics of the host population has yet to be specified. Let $\mu(x) > 0$ be the mortality of an infected host at a level of infection x, and $\mu_0 > 0$ the mortality of noninfected hosts. Let $b(x) \geqq 0$ be the fertility of infected hosts of level x, and $b_0 \geqq 0$ the fertility of noninfected hosts.

Finally, the acquisition (immigration) function φ is not prescribed a priori, but given implicitly as

$$\varphi = \beta f(w),$$

where f is a given function modeling the dynamics of vectors (e.g., insects transmitting the disease), β is a positive contact rate between hosts and vectors, and w is the average parasite load (w could be a more general functional of the population).

As described above, the population is governed by two equations of the form

$$u_t = (k(x)u_x)_x + (l(x)u)_x - \varphi(t)(\phi(x)u)_x - \mu(x)u, \tag{2a}$$

$$U_t = (b_0 - \mu_0)U - \varphi(t)\phi_0 U + \gamma u(t, 0) + \int_0^\infty b(x)u(t, x)\, dx. \tag{2b}$$

These two equations are coupled by the transition parameter γ in (2b) and by a boundary condition of the form

$$a_0 u(t, 0) + a_1 u_x(t, 0) + a_2 U(t) = 0. \tag{3}$$

The coefficients in the boundary condition are uniquely determined up to a common factor. If $\mu(x) \equiv 0$, $\mu_0 = 0$, $b(x) \equiv 0$, $b_0 = 0$ then a conservation law must hold:

$$\frac{d}{dt}\left(U(t) + \int_0^\infty u(t, x)\, dx \right) = 0. \tag{4}$$

If $u(t, x)$ converges sufficiently fast for $x \to \infty$ then (4) implies

$$
\begin{aligned}
&-\varphi(t)\phi_0 U + \gamma u(t,0) + \int_0^\infty [k(x)u_x + l(x)u - \varphi(t)\phi(x)u]_x \, dx \\
&\quad = -\varphi(t)\phi_0 U + \gamma u(t,0) - k(0)u_x - l(0)u + \varphi(t)\phi(0)u \equiv 0.
\end{aligned} \tag{5}
$$

Thus the desired boundary condition at $x = 0$ is

$$
k(0)u_x + l(0)u - \varphi(t)\phi(0)u - \gamma u = -\varphi(t)\phi_0 U. \tag{6}
$$

The most natural boundary condition at $x = \infty$ is

$$
u(t, \cdot) \in L^1(0, \infty). \tag{7}
$$

However the mathematical discussion of the initial value problem is simpler in an appropriate Hilbert space setting. This Hilbert space is specified later.

The initial condition reads

$$
u(0, x) = u_0(x) \quad \text{for } x \geqq 0, \qquad U(0) = U_0. \tag{8}
$$

The transmission law is given by $\varphi = \beta f(w)$, where

$$
w = \frac{\int_0^\infty x u(t, x)\, dx}{U(t) + \int_0^\infty u(t, x)\, dx}. \tag{9}
$$

As announced earlier, we shall restrict our attention to the linear part of the problem, i.e., we consider (2), (6) for a constant φ.

3. Transformation to normal form. For the following arguments we can neglect the biological interpretation of the coefficient functions and introduce a condensed notation by

$$
\begin{aligned}
&\tilde{l} = l - \varphi\phi, \\
&\tilde{b}_0 = b_0 - \mu_0 - \varphi\phi_0, \\
&\kappa = \varphi\phi_0 / k(0), \\
&\sigma = [\gamma - l(0) + \varphi\phi(0)]/k(0).
\end{aligned} \tag{10}
$$

Then the equations read

$$
u_t = (k(x)u_x)_x + (\tilde{l}(x)u)_x - \mu(x)u, \tag{11}
$$

$$
U_t = \tilde{b}_0 U + \int_0^\infty b(x)u(t, x)\, dx + \gamma u(t, 0),
$$

$$
u_x(t, 0) = \sigma u(t, 0) - \kappa U(t). \tag{12}
$$

This system describes an evolution equation in a product space $X = Y \times \mathbb{R}$, where Y is some function space on $\mathbb{R}_+$. The right-hand side describes diffusion on a set $P \cup [0, \infty)$, where P is a compartment. It also contains a Lotka birth law. The diffusion part can be interpreted as a self-adjoint operator in various ways, the Lotka part (as its finite-dimensional analogon, the Leslie matrix) is definitely nonself-adjoint and must be interpreted as a nonself-adjoint perturbation. This concept can be carried through in various ways depending on how the right-hand side is split into a main part and a perturbation. In the following we apply a Liouville transformation to achieve the normal form for the diffusion part.

The Liouville transformation is defined as follows. Define the function

$$
\beta(x) = \int_0^x k(s)^{-1/2}\, ds; \tag{13}
$$

then the new independent variable $y = \beta(x)$. Let $x = \alpha(y)$ be the inverse transformation. Then define the function $\rho = \rho(y)$ by

$$\rho(y) = \exp\left\{-\frac{1}{4}\int_0^x (k_x(s) + 2\tilde{l}(s))k(s)^{-1}\,ds\right\}. \tag{14}$$

Then the new dependent variables are introduced by

$$u(t, x) = \rho(y)w(t, y), \qquad U = \frac{\tilde{\sigma}}{\tilde{\kappa}}W.$$

The function w satisfies

$$w_t(t, y) = w_{yy}(t, y) - p(y)w(t, y), \tag{15}$$

where the function p is defined by

$$p(y) = \frac{1}{4}k_{xx}(x) - \frac{1}{16}\frac{k_x^2(x)}{k(x)} - \frac{1}{2}\tilde{l}_x(x) + \frac{1}{4}\frac{\tilde{l}^2(x)}{k(x)} + \mu(x). \tag{16}$$

Then functions w and u satisfy the equation

$$\int_0^\infty w^2(t, y)\,dy = \int_0^\infty m(x)u^2(t, x)\,dx, \tag{17}$$

with

$$m(x) = k(x)^{-1/2}\exp\left(\frac{1}{2}\int_0^x \frac{k_x(s) + 2\tilde{l}(s)}{k(s)}\,ds\right). \tag{18}$$

The proof is well known and will be reproduced only for the convenience of the reader. In

$$\begin{aligned} u_t &= (ku_x)_x + (\tilde{l}u)_x - \mu u \\ &= ku_{xx} + k_x u_x + \tilde{l}u_x + \tilde{l}_x u - \mu u \end{aligned}$$

insert

$$u_x = u_y\beta_x, \qquad u_{xx} = u_{yy}\beta_x^2 + u_y\beta_{xx}$$

to obtain

$$u_t = k\beta_x^2 u_{yy} + (k_x\beta_x + k\beta_{xx} + \tilde{l}\beta_x)u_y + (\tilde{l}_x - \mu)u,$$

and then

$$u = \rho w, \quad u_y = \rho w_y + \rho_y w, \quad u_{yy} = \rho w_{yy} + 2\rho_y w_y + \rho_{yy} w$$

which gives

$$\begin{aligned} w_t = k\beta_x^2 w_{yy} &+ \left(k\beta_x^2 2\frac{\rho_y}{\rho} + k_x\beta_x + k\beta_{xx} + \tilde{l}\beta_x\right)w_y \\ &+ \left(k\beta_x^2\frac{\rho_{yy}}{\rho} + k_x\beta_x\frac{\rho_y}{\rho} + k\beta_{xx}\frac{\rho_y}{\rho} + \tilde{l}\beta_x\frac{\rho_y}{\rho} + \tilde{l}_x - \mu\right)w. \end{aligned}$$

The special choice of β produces $\beta_x = k^{-1/2}$, $\beta_{xx} = -\frac{1}{2}k^{-3/2}k_x$, and thus

$$\begin{aligned} w_t = w_{yy} &+ \left(2\frac{\rho_y}{\rho} + \frac{1}{2}k^{-1/2}k_x + \tilde{l}k^{-1/2}\right)w_y \\ &+ \left(\frac{\rho_{yy}}{\rho} + \frac{1}{2}k^{-1/2}k_x\frac{\rho_y}{\rho} + \tilde{l}k^{-1/2}\frac{\rho_y}{\rho} + \tilde{l}_x - \mu\right)w. \end{aligned}$$

The special choice of ρ annihilates the coefficient of the first derivative and produces the potential p given in (16).

Define the new coefficients

$$\tilde{b}(y) = b(\alpha(y)\rho(y))\sqrt{k(\alpha(y))}, \tag{19}$$

$$\tilde{\sigma} = \sigma\sqrt{k(0)} + \frac{1}{4}\frac{k_x(0)}{\sqrt{k(0)}} + \frac{1}{2}\frac{\tilde{l}(0)}{\sqrt{k(0)}}, \tag{20}$$

$$\tilde{\kappa} = \kappa\sqrt{k(0)}. \tag{21}$$

We write again u instead of w. Then equations (2) and (6) assume the form

$$u_t = u_{yy} - p(y)u, \tag{22a}$$

$$\begin{aligned} U_t &= \tilde{b}_0 U + \frac{\tilde{\kappa}}{\tilde{\sigma}}\int_0^\infty \tilde{b}(y)u(t,y)\,dy + \frac{\tilde{\kappa}}{\tilde{\sigma}}\gamma u(t,0) \\ &= \tilde{\sigma}(u(t,0) - U(t)) + (\tilde{b}_0 + \tilde{\sigma})U(t) \\ &\quad + \frac{\tilde{\kappa}}{\tilde{\sigma}}\int_0^\infty \tilde{b}(y)u(t,y)\,dy + \left(\frac{\tilde{\kappa}}{\tilde{\sigma}}\gamma - \tilde{\sigma}\right)u(t,0), \end{aligned} \tag{22b}$$

$$u_y(t,0) = \tilde{\sigma}(u(t,0) - U(t)). \tag{23}$$

We have shown the following proposition.

PROPOSITION 1. *There is a one-to-one correspondence between the solutions of* (11), (12) *and* (22), (23). *If the solution of* (11), (12) *is in* $L^2_m(0,\infty)$, *then the solution of* (22), (23) *is in* $L^2(0,\infty)$.

4. Existence of solutions. Define the Hilbert space $X = L^2(0,\infty)\times\mathbb{C}$ with the inner product

$$\left\langle \begin{pmatrix} u \\ U \end{pmatrix}, \begin{pmatrix} v \\ V \end{pmatrix} \right\rangle = \int_0^\infty u\bar{v}\,dy + U\bar{V}. \tag{24}$$

In X define the operator L by

$$L\begin{pmatrix} u \\ U \end{pmatrix} = \begin{pmatrix} u'' - pu \\ \tilde{\sigma}u(0) - \tilde{\sigma}U \end{pmatrix} \tag{25}$$

with domain

$$D(L) = \left\{ \begin{pmatrix} u \\ U \end{pmatrix} \in X;\ u \in C_0^\infty(\mathbb{R}_+),\ u'(0) = \tilde{\sigma}(u(0) - U) \right\}. \tag{26}$$

Of course $D(L)$ is dense in X. We easily verify that L is symmetric and that the following identity holds:

$$\left\langle L\begin{pmatrix} u \\ U \end{pmatrix}, \begin{pmatrix} u \\ U \end{pmatrix} \right\rangle = -\tilde{\sigma}|u(0) - U|^2 - \int_0^\infty |u'|^2\,dy - \int_0^\infty p|u|^2\,dy. \tag{27}$$

Here is the appropriate place to introduce the qualitative properties that we require for the coefficients. We assume that there is a real constant c_0 such that

$$-k''(y) \leqq c_0, \quad k'^2(y)/k(y) \leqq c_0, \quad \tilde{l}'(y) \leqq c_0 \quad \text{for all } y \geqq 0. \tag{28}$$

Then there is a constant c such that

$$p(y) \geqq c \quad \text{for all } y \geqq 0. \tag{29}$$

We also assume that $\tilde{\sigma}$ is nonnegative, i.e.,

$$\tilde{\sigma}\sqrt{k(0)} = \gamma - \tfrac{1}{2}(l(0) - \varphi\phi(0)) + \tfrac{1}{4}k'(0) \geqq 0. \tag{30}$$

Hence the operator L is bounded above with constant c. Finally we assume $\tilde{b} \in L^2(0, \infty)$.

PROPOSITION 2. *The operator L is essentially self-adjoint.*

In the proof we use the following results ([9, II, 182, 184]).

THEOREM A. *Let H be a complex Hilbert space and A with $D(A)$ a symmetric operator that is strictly positive definite. Then the following are equivalent*:

(i) *A is essentially self-adjoint*;

(ii) $\operatorname{im}(A)$ *is dense*;

(iii) $\ker(A^*) = \{0\}$;

(iv) *A has exactly one self-adjoint extension bounded below.*

THEOREM B. *Let $u \in L^1(\mathbb{R}^n)_{\mathrm{loc}}$ such that $\Delta u \in L^1(\mathbb{R}^n)_{\mathrm{loc}}$ in the distributional sense. Define* $\operatorname{sgn} u \in L^\infty$ *by*

$$(\operatorname{sgn} u)(x) = \begin{cases} 0 & \text{if } u(x) = 0, \\ \overline{u(x)} \cdot |u(x)|^{-1} & \text{if } u(x) \neq 0. \end{cases}$$

Then, in the distributional sense,

$$\Delta|u| \geqq \operatorname{Re}[(\operatorname{sgn} u)\, \Delta u].$$

Proof of Proposition 2. The expressions φ, ϕ, l used in this proof have not the same meaning as in (2).

Define the functional l on $C_0^\infty(\mathbb{R}) \times \mathbb{C}$ by

$$l(\varphi, \phi) = \varphi'(0) - \tilde{\sigma}(\varphi(0) - \phi). \tag{31}$$

Define $\tilde{L} = L - (1-c)I$, $D(\tilde{L}) = D(L)$. Then $\tilde{L}$ is strictly negative. In view of Theorem A it is sufficient to show $\ker(\tilde{L}^*) = \{0\}$. Thus suppose $(u, U) \in X$ such that $\langle \tilde{L}(\varphi, \phi), (u, U)\rangle_{\mathbb{R}_+} = 0$ for all $(\varphi, \phi) \in D(\tilde{L})$.

Choose any fixed $(\varphi_1, \phi_1) \in C_0^\infty(\mathbb{R}) \times \mathbb{C}$ such that $l(\varphi_1, \phi_1) = 1$. For $(\varphi, \phi) \in C_0^\infty(\mathbb{R}) \times \mathbb{C}$ define

$$(\psi, \Psi) = (\varphi, \phi) - l(\varphi, \phi) \cdot (\varphi_1, \phi_1). \tag{32}$$

Then, trivially

$$(\varphi, \phi) = (\psi, \Psi) + l(\varphi, \phi) \cdot (\varphi_1, \phi_1),$$

and

$$(\psi, \Psi)|_{\mathbb{R}_+} \in D(\tilde{L}).$$

Hence

$$\langle \tilde{L}(\psi, \Psi)|_{\mathbb{R}_+}, (u, U)\rangle_{\mathbb{R}_+} = 0.$$

On the other hand,

$$\langle \tilde{L}(\psi, \Psi)_{\mathbb{R}_+}, (u, U)\rangle_{\mathbb{R}_+} = l(\varphi, \phi) \cdot \xi, \tag{33}$$

where

$$\xi = \langle \tilde{L}(\varphi_1, \phi_1)|_{\mathbb{R}_+}, (u, U)\rangle_{\mathbb{R}_+} \tag{34}$$

does not depend on (φ, ϕ).

Now extend the functions u and p to all of $\mathbb{R}$ by defining $u=0$ and $p=c-1$ for $x<0$. Then

$$\langle\varphi''-(p+1-c)\varphi, u\rangle_{\mathbb{R}}+\tilde{\sigma}\varphi(0)\bar{U}-(\tilde{\sigma}+1-c)\phi\bar{U}=(\varphi'(0)-\tilde{\sigma}\varphi(0)+\tilde{\sigma}\phi)\xi. \tag{35}$$

Now derivatives are taken in the sense of distributions. Then (35) can be written in the form

$$\langle\varphi, u''-(p+1-c)u+\tilde{\sigma}\bar{U}\delta_0+\xi\delta_0'+\tilde{\sigma}\xi\delta_0\rangle_{\mathbb{R}}=\phi((\tilde{\sigma}+1-c)\bar{U}+\tilde{\sigma}\xi) \tag{36}$$

where, as usual, $\delta_0\varphi=\varphi(0)$, $\delta_0'\varphi=-\varphi'(0)$.

Hence, in the sense of distributions, the following two equations hold:

$$u''-(p+1-c)u+\tilde{\sigma}U\delta_0+\xi\delta_0'+\tilde{\sigma}\xi\delta_0=0, \tag{37a}$$

$$(\tilde{\sigma}+1-c)\bar{U}+\tilde{\sigma}\xi=0, \tag{37b}$$

or, replacing $\bar{U}$ from the second equation,

$$u''-(p+1-c)u+\xi\delta_0'+\frac{\tilde{\sigma}(1-c)}{\tilde{\sigma}+1+c}\xi\delta_0=0. \tag{38}$$

Now define

$$\nu(x)=\max(x, 0) \tag{39}$$

and define its derivatives (in the sense of distributions) $H(x)=\nu'(x)$ (the Heaviside function), $H'(x)=\nu''(x)=\delta_0(x)$, and $H''(x)=\delta_0'(x)$.

Then (38) reads

$$u''-(p+1-c)u+\xi H''+\hat{\sigma}\xi\nu''=0, \tag{40}$$

where

$$\hat{\sigma}=\frac{\tilde{\sigma}(1-c)}{\tilde{\sigma}+1-c} \tag{41}$$

or

$$(u+h)''-(p+1-c)u=0, \tag{42}$$

where the function h is defined as

$$h(x)=\xi H(x)+\hat{\sigma}\xi\nu(x)-\xi(x\hat{\sigma}+1). \tag{43}$$

Then $h(x)=0$ for $x\geqq 0$.

Since the functions p and u are both in $L^2(\mathbb{R})_{\text{loc}}$, the function $\Delta(u+h)=(u+h)''$ is in $L^1(\mathbb{R})_{\text{loc}}$. Hence Kato's inequality (Theorem B) can be applied:

$$\begin{aligned}0&=\operatorname{sgn}(\overline{u+h})[(u+h)''-(p+1-c)u]\\&\leqq|u+h|''-\operatorname{sgn}(\overline{u+h})(p+1-c)(u+h)+\operatorname{sgn}(\overline{u+h})(p+1-c)h,\end{aligned}$$

since $h=0$ for $x\geqq 0$ and $p=c-1$ for $x\leqq 0$,

$$\begin{aligned}0&\leqq|u+h|''-(p+1-c)|u+h|\\&\leqq|u+h|''-|u+h|,\end{aligned}$$

i.e.,

$$(I-\Delta)|u+h|\leqq 0. \tag{44}$$

Since $u \in L^2(\mathbb{R})$ and h is of polynomial growth the function $|u+h|$ is a tempered distribution (i.e., a continuous linear functional on the Schwarz space $\mathcal{S}(\mathbb{R})$). The map $I-\Delta$ is a bijection on $\mathcal{S}(\mathbb{R})$ with inverse given by

$$(I-\Delta)^{-1}f(x)=\frac{1}{2}\int_{-\infty}^{\infty} e^{-|x-y|}f(y)\,dy \quad \text{for } f\in\mathcal{S}(\mathbb{R}).$$

It follows that $(I-\Delta)^{-1}$ is positive on $\mathcal{S}(\mathbb{R})'$; hence (44) implies $|u+h|\leqq 0$. Thus $u+h=0$. On the other hand, by definition, $h=0$ for $x\geqq 0$ and $u=0$ for $x<0$, and thus $u=0$, $h=0$. Consequently also $\xi=0$ and $U=0$. Thus $(u, U)=0$ as desired. Proposition 2 is proved.

By definition the closure $\bar{L}$ of L is a self-adjoint operator. Trivially also $\bar{L}$ has the upper bound zero. By the spectral mapping theory $\|(\mu I-\bar{L})^{-1}\|\leqq \mu^{-1}$ for $\mu>0$. In view of the Hille-Yosida Theorem the operator $\bar{L}$ is a generator of a contractive C_0-*semigroup* $T(t)$, $t\geqq 0$.

From [9, Thm. X. 52], it follows immediately that the semigroup is holomorphic of angle $\pi/2$.

Define the operator L_1 with $D(L_1)=D(L)$ and

$$L_1\begin{pmatrix} u \\ U \end{pmatrix}=\begin{pmatrix} 0 \\ (\tilde{b}_0+\tilde{\sigma})U+\frac{\tilde{\kappa}}{\tilde{\sigma}}\int_0^{\infty}\tilde{b}(y)u(y)\,dy+\left(\frac{\tilde{\kappa}}{\tilde{\sigma}}\gamma-\tilde{\sigma}\right)u(0) \end{pmatrix}. \tag{45}$$

PROPOSITION 3. *The operator L_1 is L-bounded with L-bound* 0.

Proof (cf. [6, IV., Ex. 1.8]). Let $u\in C^2(\mathbb{R}_+)$. Then for any x, $y\geqq 0$ and $r>0$ the following equality trivially holds:

$$u(x)=\frac{r+y-x}{r}u(y)+\frac{x-y}{r}u(y+r)+\frac{1}{r}\int_y^{y+r}\int_y^x\int_z^s u''(\tau)\,d\tau\,ds\,dz. \tag{46}$$

Choose a, b such that $b\geqq a\geqq 0$, $b-a=r>0$. Assume $a\leqq x$, $y\leqq b$. Then $|y-x|\leqq r$ and thus

$$\begin{aligned} |u(x)| &\leqq 2|u(y)|+|u(y+r)|+\frac{1}{r}\left|\int_y^{y+r}\int_y^x\int_z^s u''(\tau)\,d\tau\,ds\,dz\right| \\ &\leqq 2|u(y)|+|u(y+r)|+\frac{1}{r}\left|\int_y^{y+r}\int_y^x\int_z^s [u''(\tau)-p(\tau)u(\tau)]\,d\tau\,ds\,dz\right| \\ &\quad +\frac{1}{r}\left|\int_y^{y+r}\int_y^x\int_z^s p(\tau)u(\tau)\,d\tau\,ds\,dz\right|. \end{aligned} \tag{47}$$

Now integrate over y from a to b and apply the Cauchy-Schwarz inequality

$$r|u(x)|\leqq 3\sqrt{r}\|u\|+r^2\sqrt{2r}\|u''-pu\|+r^2\left(\int_a^{b+r}p^2(s)\,ds\right)^{1/2}\|u\|. \tag{48}$$

Hence

$$|u(x)|\leqq\left[\frac{3}{\sqrt{r}}+r\left(\int_a^{b+r}p^2(s)\,ds\right)^{1/2}\right]\|u\|+r\sqrt{2r}\|u''-pu\|.$$

In particular, for $x=0$

$$|u(0)|\leqq c_1\|u\|+c_2\|u''-pu\|, \tag{49}$$

where

$$c_1 = \frac{3}{\sqrt{r}} + r\left(\int_0^{2r} p^2(s)\,ds\right)^{1/2}, \qquad c_2 = r^2\sqrt{2r}. \tag{50}$$

Of course c_2 can be made arbitrarily small.

Since $\tilde{b} \in L^2(0,\infty)$, we have

$$\left|\int_0^{\infty} \tilde{b}(x)u(x)\,dx\right| \leqq \|\tilde{b}\| \cdot \|u\|. \tag{51}$$

Hence for the operator L_1 the following inequality holds:

$$\left\|L_1\begin{pmatrix} u \\ U \end{pmatrix}\right\| \leqq (\tilde{b}_0 + \tilde{\sigma})|U| + \left|\frac{\tilde{\kappa}}{\tilde{\sigma}}\gamma - \tilde{\sigma}\right|\{c_1\|u\| + c_2\|u'' - pu\|\} + \left|\frac{\tilde{\kappa}}{\tilde{\sigma}}\right| \cdot \|\tilde{b}\| \cdot \|u\|. \tag{52}$$

Define

$$\begin{aligned} \alpha &= 2\max\left\{b_0 + \tilde{\sigma}, \left|\frac{\tilde{\kappa}}{\tilde{\sigma}}\gamma - \tilde{\sigma}\right| c_1 + \left|\frac{\tilde{\kappa}}{\tilde{\sigma}}\right| \cdot \|\tilde{b}\|\right\}, \\ \beta &= 2\left|\frac{\tilde{\kappa}}{\tilde{\sigma}} - \tilde{\sigma}\right| c_2. \end{aligned} \tag{53}$$

Then

$$\left\|L_1\begin{pmatrix} u \\ U \end{pmatrix}\right\| \leqq \alpha\left\|\begin{pmatrix} u \\ U \end{pmatrix}\right\| + \beta\left\|L\begin{pmatrix} u \\ U \end{pmatrix}\right\|. \tag{54}$$

Since β can be made arbitrarily small, the operator L_1 is L-bounded with L-bound 0.

Now we can define the extension $\bar{L}_1$ of L_1 from $D(L)$ to $D(\bar{L})$. Assume $(u, U) \in D(\bar{L})$ and let $(u_n, U_n) \in D(L)$ be a sequence with $(u_n, U_n) \to (u, U)$. Then, in view of (54), the sequence $L_1(u_n, U_n)$ is a Cauchy sequence. Its limit is defined as $\bar{L}_1(u, U)$. Trivially this definition is independent of the choice of the sequence. Hence the operator $\bar{L}_1$ with domain $D(\bar{L})$ is defined. Finally define the operator A with $D(A) = D(\bar{L})$ by $A = \bar{L} + \bar{L}_1$.

For the last step of the existence proof we use the following two theorems ([6, p. 497 ff], [8, p. 8]; see also [7]).

THEOREM C. *Let $\bar{L}$ be the generator of a quasibounded holomorphic semigroup and let $\bar{L}_1$ be an $\bar{L}$-bounded operator with $\bar{L}$-bound 0. Then $\bar{L} + \bar{L}_1$ is the generator of a quasibounded holomorphic semigroup.*

THEOREM D. *Let A with domain $D(A)$ be the generator of a holomorphic semigroup $T(t)$ on a Banach space X. Then for every initial value $u_0 \in X$ the Cauchy problem*

$$\frac{d}{dt}u(t) = Au(t), \qquad u(0) = u_0$$

has a unique solution $u(t)$ which is continuous for $t \geqq 0$, continuously differentiable and $u(t) \in D(A)$ for $t > 0$. This solution is given by $u(t) = T(t)u_0$.

If $u_0 \in D(A)$ then $u \in C^1(\mathbb{R}_+, X)$.

The immediate consequence of these theorems and of Propositions 1, 2, 3 is the following theorem.

THEOREM 4. *The Cauchy problem* (22), (23) *has a unique solution for initial data in $L^2(0,\infty) \times \mathbb{C}$, and the Cauchy problem* (2), (3), *with φ constant, has a unique solution for initial data in $L^2_m(0,\infty) \times \mathbb{C}$, with weight function m given by* (18).

REFERENCES

[1] R. M. ANDERSON AND R. M. MAY, *Population dynamics of infectious diseases: Part* I, Nature, 280 (1979), pp. 361-367.

[2] R. M. MAY AND R. M. ANDERSON, *Population dynamics of infectious diseases: Part* II, Nature, 280 (1979), pp. 455-461.

[3] K. P. HADELER AND K. DIETZ, *Nonlinear hyperbolic partial differential equations for the dynamics of parasite populations*, Comp. Math. Appl., 9 (1983), pp. 415-430.

[4] ———, *Population dynamics of killing parasites which reproduce in the host*, J. Math. Biol., 21 (1984), pp. 45-65.

[5] K. P. HADELER, *Vector-transmitted diseases in structured populations*, Proc. Workshop on Dynamical Systems and Environmental Models, Wartburg, March 1986.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, Heidelberg, 1966.

[7] S. G. KREIN, *Linear differential equations in Banach space*, in Translations of Mathematical Monographs, Vol. 29, American Mathematical Society, Providence, RI, 1971.

[8] A. PAZY, *Semigroups of Linear Operators and Applications to* P.D.E., Springer-Verlag, Berlin, New York, Heidelberg, 1983.

[9] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics, Vol.* I, II, Academic Press, New York, 1972 and 1975.

[10] R. WALDSTÄTTER, *Ein Populationsmodell parasitärer Erkrankungen mit stetiger Parasitenlast*, Diplomarbeit, Mathematische Fakultät, Universität Tübingen, 1986.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

008

EXISTENCE OF SOLUTIONS OF THE SIMILARITY EQUATIONS FOR FLOATING RECTANGULAR CAVITIES AND DISKS*

CHUNQING LU†¶, NICHOLAS D. KAZARINOFF†,
J. BRYCE McLEOD‡, AND WILLIAM C. TROY§

Abstract. The differential equation $f'''+Q[Aff''-(f')^2]=\beta$ $(0\leqq A<\infty,\ Q>0,\ \beta$ real$)$ $(' = d/dx)$ for $0\leqq x\leqq 1$ with boundary conditions $f(0)=f(1)=0$, $f''(1)=f''(0)+1=0$ is considered. Existence of at least one solution of this two-point boundary-value problem is proved under various hypotheses, and some qualitative properties of this solution are established. The main tools used are shooting arguments and the Schauder fixed point theorem.

Key words. existence, similarity equations, nonlinear two-point boundary-value problem

AMS(MOS) subject classifications. 34B10, 34A, 76

1. Introduction.

We consider the differential equation

$$f'''+Q[Aff''-(f')^2]=\beta \qquad (0\leqq A<\infty,\ Q>0,\ \beta \text{ real}) \tag{1}$$

$(' = d/dx)$ for $0\leqq x\leqq 1$ subject to the boundary conditions

$$f(0)=f(1)=f''(1)=f''(0)+1=0. \tag{2}$$

In this paper we prove existence of at least one solution of the two-point boundary-value problem (TPBVP) (1)-(2) under various hypotheses, and we establish some qualitative properties of this solution. The results are as follows.

THEOREM 1. *For each given $\beta\in(0,1)$ and each $A\geqq 0$ there exists at least one $Q>0$ such that* (1) *subject to the boundary conditions* (2) *having at least one solution.*

THEOREM 2. *If $A=2$, then for each $Q>0$ there exists at least one real number β, with $1>\beta>1-\frac{1}{4}Q$, such that the problem* (1)-(2) *has a nonnegative, convex solution.*

THEOREM 3. *If $A=1$, then for each $Q>0$ there exists at least one real number β with $1>\beta>1-\frac{1}{4}Q$ such that the problem* (1)-(2) *has a nonnegative, convex solution.*

Remarks. If $Q=0$, then for any A (1)-(2) has the unique polynomial solution $x(x-1)(x-2)/6$ corresponding to $\beta=1$. The conclusion of Theorem 1 holds for any $A>0$, although for the model described below $A\geqq 1$.

The TPBVP (1)-(2) arises from a reduction by similarity of the boundary-layer formulation of the Navier-Stokes equations for the distributions of velocity in a low Prandtl number fluid zone in the shape of either a floating rectangular slot or a floating circular disk [2], [3]. Here "floating" means that two opposite surfaces of the rectangular cavity are free surfaces and that the two opposite surfaces of the disk are free. The flow in the low Prandtl number fluid (liquid metal or silicon) is contained by the lateral solid surfaces and surface tension. A temperature gradient caused by radiation heating is assumed to exist on the free surfaces. This gradient from the hot midline (center for the disk) to the cold solid walls (wall) drives the flow. The floating zones are assumed to be in a microgravity environment, as on the space shuttle, so that the force of gravity may be neglected. The physical coordinates (x, y) and velocities (u, v) for the slot are

* Received by the editors April 6, 1987; accepted for publication (in revised form) October 10, 1987.
† Department of Mathematics, State University of New York, Buffalo, New York 14214-3093.
‡ Wadham College, Oxford University, Oxford, England.
§ Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.
¶ Present address, Computing Center, Academia Sinica, P.O. Box, 2719, Beijing, People's Republic of China.

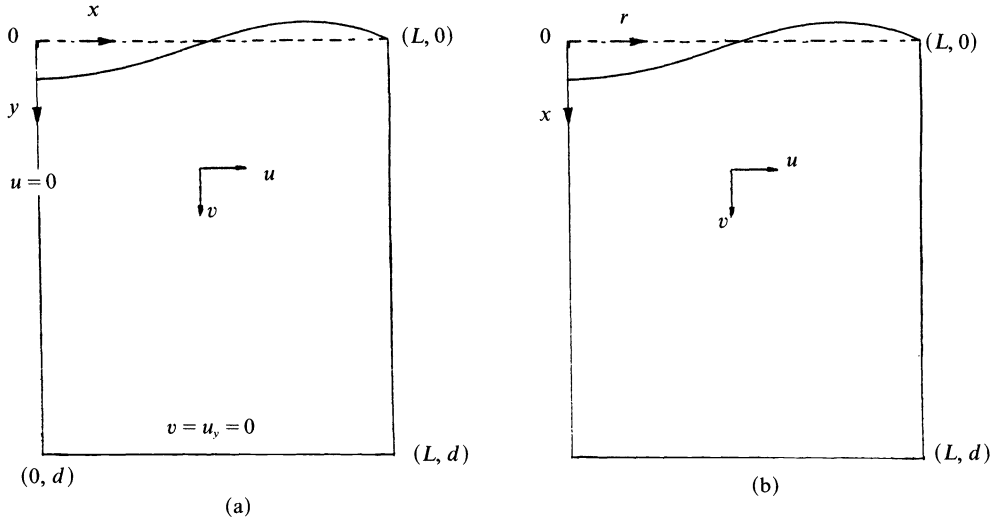


FIG. 1. (a) *The coordinates* (x, y) *and velocity components* (u, v) *for a floating slot. One quarter of a cross-section is shown.* (b) *The coordinates* (x, y) *and velocity components for a floating disk. One quarter of a section made by a bisecting plane is shown. The surface deflection is exaggerated in both figures.*

shown in Fig. 1(a). For the slot, if (a) the temperature distribution varies as $c_0 + g(\eta) \times (x/L)^m$ $(\frac{1}{2} < m < 2)$, (b) the similarity variable is chosen to be $\eta = y/\delta(x)$, and (c) $(u, v) = (c_1 x^{(2m-1)/3} f'(\eta),\ (-c_2 \eta f'(\eta) + c_3 f(\eta)) x^{(m-2)/3})$ (where the c_i are suitably chosen positive constants), then, provided $\delta(x) = \text{const.} \times (x/L)^{(2-m)/3}$, the x-acceleration equation leads to (1)-(2) with $A = (m+1)/(2m-1)$ and $Q = 2(d/L)^3 \operatorname{Re} m(2m-1)/3$. Here Re is the Reynolds number. For convenience we rename η by x in this paper. For the disk, if $\eta = x/d$ and $(u, v) = (-c_1 f, c_2 r f')$, where the c_i are positive constants, then f satisfies (1)-(2) with $A = 2$. The physical coordinates (x, y) and velocities (u, v) for the disk are illustrated in Fig. 1(b).

Numerical solution [2], [3] of the TPBVP (1)-(2) has led to the bifurcation diagrams shown in Figs. 2(a) and (b) for $A = 1$ and $A = 2$, respectively. The references to two-cell flow and three-cell flow in Figs. 2(a) and (b) are related to the number of zeros of $f'(x)$ on $(0, 1)$. If f' has but one zero, then the u-component of velocity in Figs. 1(a) and (b) changes sign once in each half-zone, which corresponds to a flow with one cell in each half-zone. But if f' changes sign twice on $(0, 1)$, then $u = 0$ twice in each half-zone, and there are three flow cells, half of the middle one lying in each zone. We have verified only a small portion of the information represented in these diagrams. For the solutions we have found, f' always has but one zero on $(0, 1)$.

We prove Theorem 1 using a topological method introduced by Hastings in [1]. We prove Theorems 2 and 3 using the Schauder Fixed Point Theorem [4]. The proof of Theorem 1 is given in § 2, and the proofs of Theorem 2 and 3 are given in § 3.

2. Proof of Theorem 1. We divide the proof of Theorem 1 into several lemmas. In the proof, for a given $\beta \in (0, 1)$, we let $f(x; Q, \alpha)$ be the solution of (1) satisfying

$$f(0; Q, \alpha) = f''(0; Q, \alpha) + 1 = 0, \qquad f'(0; Q, \alpha) = \alpha; \tag{3}$$

we define four subsets of points (Q, α) in $R_+^2 = (0, \infty) \times (0, \infty)$ as follows:

$$\begin{aligned} S_1 &= \{(Q, \alpha) \mid f(1; Q, \alpha) > 0\}, \qquad S_2 = \{(Q, \alpha) \mid f(1; Q, \alpha) < 0\} \\ S_3 &= \{(Q, \alpha) \mid f''(1; Q, \alpha) > 0\}, \qquad S_4 = \{(Q, \alpha) \mid f''(1; Q, \alpha) < 0\}. \end{aligned} \tag{4}$$

Our proof is based on Hasting's Lemma 3 in [1] and an additional argument in [1,

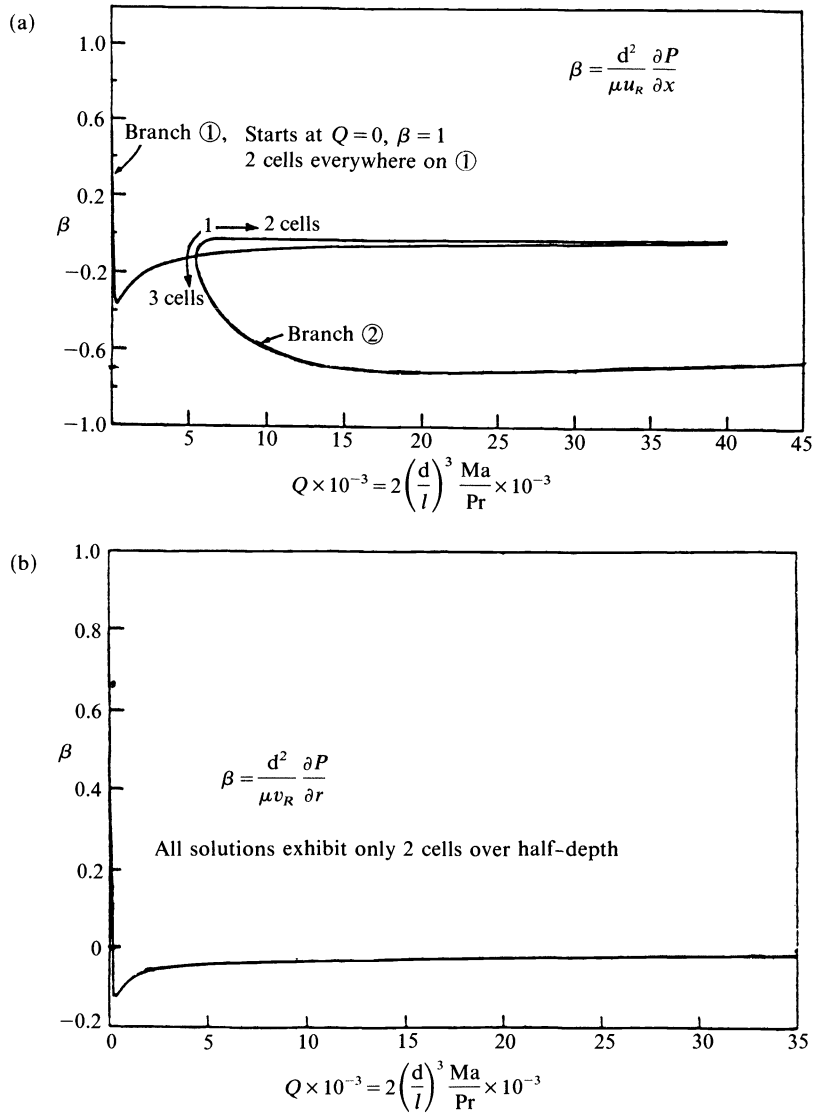


FIG 2. *Bifurcation diagrams for a floating slot (problem* (1)-(2) *with* $A = 1$*) and the floating disk (problem* (1)-(2) *with* $A = 2$*):* β *versus* $Q \times 10^{-3}$. (a) *The slot*; (b) *the disk.*

pp. 106-107]. His lemma is: Suppose S_i $(i = 1, \cdots, 4)$ are open sets of R_+^2 with $S_1 \cap S_2 = \varnothing$ and $S_3 \cap S_4 = \varnothing$. Further, suppose there are components $P_i \subset S_i$ such that $P_1 \cap P_4$, $P_1 \cap P_3$, $P_2 \cap P_3$, and $P_2 \cap P_4$ are not empty. Then $S = S_1 \cup S_2 \cup S_3 \cup S_4 \neq R_+^2$. If $P_2 \cap P_3$ is not empty, then Hastings' lemma applies. However, we are not able to determine whether or not $P_2 \cap P_3 = \varnothing$. If $P_2 \cap P_3 = \varnothing$, we follow Hastings [1, pp. 106-107] to construct and use a nonempty, unbounded, connected subset $W \subset S_2 \cup S_3$ to show that for each $\beta \in (0, 1)$ the complement of S in R_+^2 contains at least one point (Q_0, α_0) that corresponds to a solution of (1)-(2). We observe that several times in the sequel we use the same symbols P_i for a component of S_i and for a subset of that component. For brevity, we denote solutions of (1)-(3) simply by $f(x)$, suppressing their dependence on Q and α.

It is obvious that S_i $(i=1,\cdots,4)$ is open, $S_1 \cap S_2 = \varnothing$, and $S_3 \cap S_4 = \varnothing$. We first prove that S_1 is not empty and that, if $\alpha > 1$, then f is positive everywhere on $(0, 1)$.

LEMMA 1. *If $\alpha > 1$ and $0 < \beta < 1$, then for any $Q > 0$, $(Q, \alpha) \in S_1$ $(S_1 \neq \varnothing)$, $f'(x) \geqq 0$ on $[0, 1]$, and f is monotonic increasing on $(0, 1)$.*

Proof. Choose $\alpha > 1$ and $0 < \beta < 1$. Then $f(0)f''(0) \leqq 0$, and

$$f''' \geqq \beta + Q(f')^2 \geqq \beta > 0 \tag{5}$$

as long as $ff'' \leqq 0$. Integrating (5) twice and using (3), we find that

$$f'(x) \geqq \beta x^2/2 - x + \alpha > \beta x^2/2 + (-x+1) > 0. \tag{6}$$

Consequently, if $\alpha > 1$, then $f' > 0$ so long as $ff'' \leqq 0$. Thus $f'(x) \geqq 0$ for $0 \leqq x \leqq 1$. If not, $f'(x_1) < 0$ for some $x_1 \in (0, 1)$. Then there exists an x_0, the first zero of f' at which $f''(x_0) < 0$ and $f(x_0) > 0$. We claim that $ff'' \leqq 0$ on $(0, x_0)$. If not, since $f > 0$ on $(0, x_0)$ there is an $x_3 \in (0, x_0)$ with $f''(x_3) = 0$ and $f'''(x_3) < 0$, which is impossible. Hence, $ff'' \leqq 0$ on $(0, x_0)$ and $f'(x_0) = 0$, which contradicts (6). Thus, $f'(x) \geqq 0$ on $[0, 1]$, and since $f'(0) = \alpha > 1$, $f(1) > 0$. □

We next prove the following.

LEMMA 2. *Each of S_4, $S_1 \cap S_4$, and $S_2 \cap S_4$ is nonempty.*

Proof. If $Q = 0$, then $f''(1; 0, \alpha) = \beta - 1 < 0$. Let $\alpha^* = \frac{1}{2}(1 + \beta/3)$. Then, if $\alpha > \alpha^*$, $f(1; 0, \alpha) > 0$; and if $\alpha < \alpha^*$, $f(1; 0, \alpha) < 0$. By continuity of solutions of (1)-(3) in Q, the conclusions of Lemma 2 now follow.

Remark. What we have actually proved in Lemma 2 is that there exist components $P_i \subset S_i$ $(i=1,\cdots,4)$ such that $P_1 \cap P_4$ and $P_2 \cap P_4$ are not empty.

In order to show that the union of the four S_i is not all of R_+^2, we use two additional lemmas, Lemmas 3 and 4 below. Finally, as indicated in the Introduction, topological arguments (Hastings [1]) will be used to complete the proof of Theorem 1.

LEMMA 3. *The set $W_2 = \{(Q, \alpha) \mid Q > 0, \alpha > \max\{(6/Q)^{1/3}, (4/Q)^{1/2}\}\} \subset S_3$.*

Proof. By (5) and the hypothesis of the lemma,

$$f''' \geqq \beta + Q(f')^2 > \beta + Q\alpha^2/4$$

as long as $ff'' \leqq 0$ and $f'(x) > \alpha/2$. Then integrations give

$$f''(x) > (\beta + Q\alpha^2/4)x - 1 \tag{7}$$

and

$$f'(x) > \tfrac{1}{2}(\beta + Q\alpha^2/4)x - x + \alpha. \tag{8}$$

Since $\alpha > (6/Q)^{1/3}$, $\frac{1}{2}(\beta + Q\alpha^2/4)x - x + \alpha > 2\alpha/3$ for $x \in [0, 1]$. If $f(x) \geqq 0$ and $f''(x) \leqq 0$ for $x \in [0, 1]$, then $f'(x) > \alpha/2$ for $x \in [0, 1]$. Otherwise there must be an interval $[0, x^*)$ such that $f'(x) > \alpha/2$ on $[0, x^*)$ and $f'(x^*) = \alpha/2$, which is impossible since $\alpha > (6/Q)^{1/3}$, which implies $f'(x) \geqq 2\alpha/3$ on $[0, x^*)$. By (7), however, if $f(x) \geqq 0$ and $f''(x) \leqq 0$ for $x \in [0, 1]$, $f''(1) > (\beta + Q\alpha^2/4) - 1 > \beta > 0$ since $\alpha \geqq (4/Q)^{1/2} > 0$, which is a contradiction. Thus there must exist an $x_* \in (0, 1)$, such that $f''(x_*) > 0$, which implies that $f''(1) > 0$, because wherever $f''(x) = 0$, $f'''(x) > 0$. □

LEMMA 4. *The set $W_1 = \{(Q, \alpha)_i \mid Q > 0, 0 < \alpha < \beta/64\} \subset S_2 \cup S_3$.*

Proof. Let Q and α be given in W_1. If $f(1) < 0$, $(Q, \alpha) \in S_2$, and we are done. We shall therefore consider the following cases.

Case 1. $f(1) \geqq 0$ and there exists an $x^* \in (0, 1)$ such that $f(x^*) < 0$. Then it must be that $f''(1) > 0$. Otherwise, $f''(1) \leqq 0$ and $f(1) \geqq 0$ would imply that there would be an $x_1 \in (x^*, 1)$ such that $f''(x_1) = 0$ and $f'''(x_1) < 0$, which is a contradiction of the differential equation (1) at x_1. Therefore, $(Q, \alpha) \in S_3$.

Case 2. $f(1) \geqq 0$ and $f(x) \geqq 0$ on $[0, 1)$.

If $f''(1) \leqq 0$, then $f''(x) \leqq 0$ for all $x \in [0, 1]$; otherwise there is an x where $f''(x) = 0$ and $f'''(x) < 0$, which is a contradiction as above. Hence $f'''(x) \geqq Q[f'(x)]^2 + \beta \geqq \beta$ for all $x \in [0, 1]$. We consider two subcases.

(i) There is an $x_2 \in (0, \frac{1}{4})$ such that $f''(x_2) = -\frac{1}{4}\beta$. In this subcase,

$$\int_{x_2}^{x_2+1/2} f'''(x)\,dx = f''\left(x_2 + \frac{1}{2}\right) - f''(x_2) \geqq \frac{1}{2}\beta,$$

and hence $f''(x_2 + \frac{1}{2}) > \frac{1}{2}\beta + f''(x_2) = \frac{1}{4}\beta > 0$, which is a contradiction.

(ii) $f''(x) < -\frac{1}{4}\beta$ for all $x \in (0, \frac{1}{4}]$. In this subcase,

$$f'(x) \leqq -\tfrac{1}{4}\beta x + \alpha \qquad (x \in (0, \tfrac{1}{4}])$$

so that

$$f(x) \leqq -(\beta x^2/8) + \alpha x \leqq x(-\beta x/8 + \beta/64).$$

Hence $f(\frac{1}{4}) < 0$, which contradicts the main hypothesis that $f(x) \geqq 0$ on $[0, 1)$. Therefore $f''(1) > 0$, and $(Q, \alpha) \in S_3$.

Case 3. $f(1) < 0$. In this case $(Q, \alpha) \in S_2$. □

An important consequence of Lemma 3 is that S_3 has a nonempty component P_3. Moreover, there is a component P_1 of S_1 such that $P_1 \cap P_3 \neq \varnothing$. To see this, consider the quarter-plane

$$W_0 = \{(Q, \alpha) \mid Q > Q_0, \alpha > 1\}.$$

By Lemma 1, $W_0 \subset S_1$. But $S_1 \cap S_2 = \varnothing$. By Lemma 4, W_0 must contain points in a component P_3 of S_3, and $P_1 \cap P_3 \neq \varnothing$.

In like manner, by Lemma 4, we conclude that W_1 contains at least a subset of a component P_2 of S_2; hence by Lemma 2 $P_2 \cap P_4 \neq \varnothing$.

Using Lemmas 1–4 and the last paragraph, we can sketch portions of the four sets P_i as shown in Fig. 3. If we now define $W = W_1 \cup W_2$, an argument essentially identical to the one in [1, pp. 106–107] completes the proof of Theorem 1. (The reader

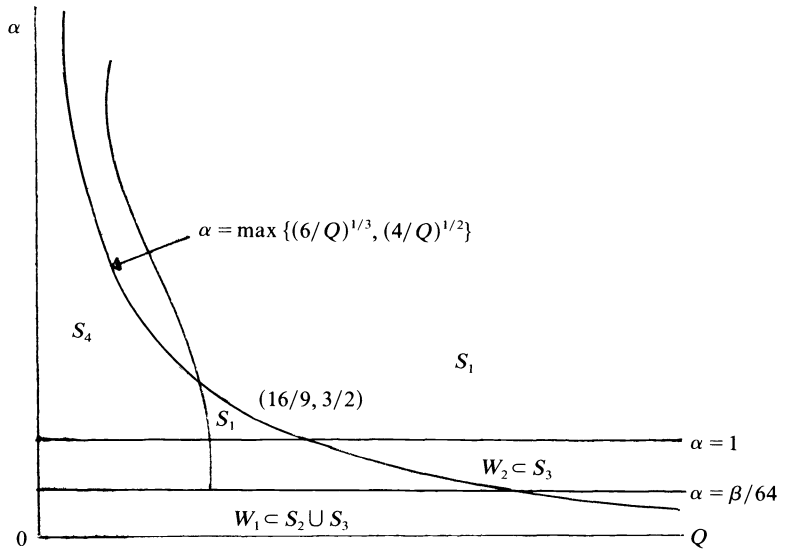


FIG 3. *The sets* S_1, S_3, S_4, *and* $S_2 \cup S_3$.

should identify our W, P_1, P_2, P_3, P_4, S_1, S_2, S_3, and S_4 with Hastings' W, R_1, Q_1, Q_2, R_2, T_1, S_1, S_2, and T_4, respectively.)

3. Proofs of Theorems 2 and 3. We begin with the proof of Theorem 2. Let $Q>0$ be given and fixed. It is sufficient to prove

$$f^{(4)}(x)+2Qff'''=0 \tag{9}$$

has a solution satisfying the boundary conditions (2). Let

$$D=\{f\,|\,f\in C^2[0,1], f(0)=f(1)=0, -1=f''(0)\leqq f''(x)\leqq f''(1)=0, \text{ and } 0\leqq f(x) \text{ on } [0,1]\},$$

where $C^2[0,1]$ is the Banach space containing all real-valued functions f twice continuously differentiable on $[0,1]$ with $\|f\|=\sup(|f|)+\sup(|f'|)+\sup(|f''|)$ over $[0,1]$. Then

(i) D is a closed, bounded, convex subset of $C^2[0,1]$. If we integrate the differential inequality for $f''(x)$ twice and use $f(1)=f(0)=0$, we find that $-1\leqq -x+\alpha\leqq f'(x)\leqq\alpha$ and $0\leqq f(x)\leqq\alpha x\leqq\frac{1}{2}$, where $0<\alpha=f'(0)\leqq\frac{1}{2}$ ($\alpha\geqq 0$ since if $\alpha=f'(0)<0$, $0\leqq f(x)<0$, which is a contradiction).

(ii) D contains nonzero elements; for example, $x(x-1)(x-2)/6\in D$.

We now define a mapping T with domain D as follows. For $f\in D$, $T(f)=f^*$, where f^* satisfies

$$f^{*(iv)}+2Qff^{*'''}=0, \qquad f^*(0)=f^*(1)=f^{*''}(0)+1=f^{*''}(0)=0.$$

We will show that T is well defined, maps D into D, and is compact. Then Schauder's Fixed Point Theorem [4] will apply to T; and, for each $Q>0$, the fixed point will be the desired solution of (1)-(2).

For any given $f\in D$, the unique solution and its derivatives are given by the formulas

$$f^{*'''}(x)=c\,e^{w(x)}, \tag{10}$$

$$f^{*''}(x)=c\int_0^x e^{w(s)}\,ds-1, \tag{11}$$

$$f^{*'}(x)=\int_0^x\left\{\int_0^v c\,e^{w(s)}\,ds-1\right\}dv+\alpha_f, \tag{12}$$

$$f^*(x)=\int_0^x\int_0^u\left\{\int_0^v c\,e^{w(s)}\,ds-1\right\}dv\,du+\alpha_f x, \tag{13}$$

where

$$w(x)=-2Q\int_0^x f(t)\,dt \tag{14}$$

and

$$1<c=f^{*'''}(0)=\left[\int_0^1 e^{w(s)}\,ds\right]^{-1}<\frac{Q}{1-e^{-Q}} \tag{15}$$

by (10) since $\frac{1}{2}\geqq f(x)$, and $\alpha_f=f^{*'}(0)$ is uniquely determined by the condition $f^*(1)=0$, namely

$$\alpha_f=\int_0^1\int_0^u\left\{1-\int_0^v c\,e^{w(s)}\,ds\right\}dv\,du. \tag{16}$$

The above formulas show that $f^* = Tf$ is well defined and continuous for $f \in D$. Since $\frac{1}{2} \geqq f(x) \geqq 0$, $Q/[1-e^{-Q}] \geqq c = f^{*'''}(0) \geqq f^{*'''}(x) \geqq f^{*'''}(1) > 0$ by (10), (14), and (15), and $f^{*'''}$ is decreasing for x increasing. Further, $-1 \leqq f^{*''}(x) \leqq 0$, and $f^{*''}$ is increasing for x increasing. From (16) it follows that $\alpha_f \leqq \frac{1}{2}$. Integrating the inequalities for $f^{*''}$ and using the boundary conditions satisfied by f^*, it follows that $\frac{1}{2} \geqq \alpha_f x \geqq f^*(x) \geqq 0$ and $-1 \leqq -x + \alpha_f \leqq f^{*'}(x) \leqq \alpha_f \leqq \frac{1}{2}$. Hence, $f^* \in D$.

We next show that T is compact, namely, that $\overline{T(K)}$, the closure of $T(K)$, is compact for every closed, bounded subset K of D. Let K be a closed, bounded subset of D, and let $\{f_i\}$ be a sequence of functions in K with images $\{f_i^*\}$ under T. Then $|f_i^{*'''}| \leqq Q/[1-e^{-Q}]$, $|f_i^{*''}| \leqq 1$, $|f_i^{*'}| \leqq 1$, and $|f_i^*| \leqq \frac{1}{2}$ on $[0, 1]$ for each i. Therefore, the f_i^*, $f_i^{*'}$, and $f_i^{*''}$ are equicontinuous on $[0, 1]$. Hence, by the Arzela-Ascoli Theorem, there exist a subsequence $\{f_{n(i)}^*\}$ of $\{f_i\}$ and a $g \in \overline{T(K)}$ such that $\|f_{n(i)}^* - g\| \to 0$ as $i \to \infty$. Thus $\overline{T(K)}$ is compact. Hence, by Schauder's Theorem [4], T is a compact operator from D into D, and there exists a fixed point f_0 of T, which by (10)-(16) is a solution of the TPBVP (1)-(2) for the given $Q > 0$, and for which $\beta = f'''(0) - Q[f_0'(0)]^2 = f'''(1) - Q[f_0'(1)]^2$. By (15), $1 < c$; and, since $\frac{1}{2} \geqq f(x) \geqq 0$, by (10), $f'''(1) \leqq 1$. Moreover $0 < f_0'(0) < \frac{1}{2}$. Therefore, $1 > \beta > 1 - Q/4$. This completes the proof of Theorem 2.

In the proof of Theorem 3, we again use the subset D of $C^2[0, 1]$, with the usual C^2-norm, used in the proof of Theorem 2. Recall that D is closed, bounded, convex, and nonempty, and that, integrating the inequalities for $f''(x)$ and using $f(1) = f(0) = 0$, we obtain the bounds $-1 \leqq f'(x) \leqq \frac{1}{2}$ and $0 \leqq f(x) \leqq \frac{1}{2}x \leqq \frac{1}{2}$. In the case considered $A = 1$ so that the differential equation (1) is: $f''' + Q[ff'' - (f')^2] = \beta$. The map T from D into $C^2[0, 1]$ now is the following: $Tf = f^*$, where f^* is a solution of the linear TPBVP

$$(17) \qquad f^{*(iv)} + Qff^{*'''} - Qf'f^{*''} = 0, \qquad f^*(0) = f^*(1) = f^{*''}(0) + 1 = f^{*''}(1) = 0.$$

Note that $f^{*(iv)}(1) = 0$ for a solution of (17) since $f(1) = f^{*''}(1) = 0$. To show that T is well defined we use a backward shooting method to solve the related second-order TPBVP

$$(18) \qquad H'' + QfH' - Qf'H = 0, \quad H(0) = -1, \quad H(1) = 0.$$

Then we shall set $f^{*''} = H$, and integrate $f^{*''}$ twice, using $f^*(0) = f^*(1) = 0$ to determine $f^{*'}(0)$ so that f^* will lie in D.

Let $H'(1) = \lambda$. We differentiate (18), multiply both sides by the integrating factor $\exp[w(s)]$, where

$$w(s) = -Q \int_s^1 f(t)\, dt,$$

and integrate both sides from one to x, using the boundary condition $H''(1) = 0$ ($f^{*(iv)}(1) = 0$ implies $H''(1) = 0$). The result is

$$(19) \qquad H''(x)\, e^{w(s)} = Q \int_1^x f''(s) H(s)\, e^{w(s)}\, ds.$$

Therefore if $\lambda < 0$, $H(x) > 0$ for x near one, and hence $H''(x) > 0$ near one. Indeed, by (19), $H''(x) > 0$ so long as $H(x) > 0$. Thus (19) and $\lambda < 0$ imply that $H(0) > 0$. Since we want $H(0) = -1$, we proceed further. If $\lambda = 0$, then $H \equiv 0$ (the solution of the backward linear initial value problem is unique). Consequently, we try $\lambda > 0$. Then since $H(1) = 0$, $H(x) < 0$; and, by (19), $H''(x) < 0$ for x close to one. Now we conclude from (19) that $H''(x) < 0$ so long as $H(x) < 0$. Thus $H(0) < 0$. Indeed, $H''(x) < 0$ on $(0, 1)$ implies $H'(x) \geqq \lambda$ on $[0, 1]$. Integrating this inequality from zero to one and

using $H(1)=0$, we obtain $H(0)\leqq -\lambda$. Thus for $\lambda=1$, $H(0)\leqq -1$. Since H is a solution of a linear initial value problem (going backward from one), H is continuous in λ. Therefore, since for $\lambda=0$, $H(0)=0$, and for $\lambda=1$, $H(0)\leqq -1$, there must be a λ_0 on $(0,1]$ such that $H(0)=-1$, as desired. Moreover, for $H'(1)=\lambda_0$, $H'(x)\geqq \lambda_0>0$, since $H''(x)<0$ on $(0,1)$. Thus $H''(x)$, $-H'(x)$, and $H(x)$ are negative on $(0,1)$. Therefore, $H(x)=f^{*''}(x)$ monotonically increases on $[0,1]$ and $-1\leqq H(x)=f^{*''}(x)\leqq 0$ for $x\in[0,1]$. By repeated integration of (19) from one to x, we can choose $f^*(1)=0$ and $f^{*'}(1)$ so that $f^*(0)=0$. Then, integrating $-1\leqq f^{*''}\leqq 0$, we obtain the bounds on $f^{*'}$ and f^* required for f^* to lie in D. Since the TPBVP (18) is linear, the function $H=f^{*''}$ satisfying (18) for $\lambda=\lambda_0$ is unique and depends continuously on f. It follows that $f^*=Tf$ is unique and depends continuously on f. Consequently, T is a continuous map from D into D. That T is compact would follow from uniform bounds on $f^{*'''}$, $f^{*''}$, $f^{*'}$, and f^*. We have established these bounds except for a uniform bound for $|f^{*'''}(x)|$. We have a lower bound for $H'=f^{*'''}$, namely $H'(1)=\lambda_0>0$.

The parameter $Q>0$ is given and fixed. Differentiating (18), we see that $H'''+QfH''-Qf''H=0$; and hence $H'''=Qf''H-QfH\geqq 0$. Therefore, $0\geqq H''(x)\geqq H''(0)$ on $[0,1]$. From (18) we conclude that $H''(0)=-Qf'(0)\geqq -Q/2$; hence $0\geqq H''\geqq -\frac{1}{2}Q$, and $-\frac{1}{2}Qx+c\leqq H'\leqq c$, where $c=H'(0)=f^{*'''}(0)$. Now, $H(0)=-1$ implies that $-\frac{1}{4}Qx^2+cx-1\leqq H\leqq cx-1$. Applying $H(1)=0$, we find that $-\frac{1}{4}Q+c-1\leqq 0$ or $1+c\leqq\frac{1}{4}Q$, and $c-1\geqq 0$. Thus, since $1\geqq\lambda_0>0$, $|H'|=|f^{*'''}|\leqq 1+\frac{1}{4}Q$.

We now can assert that T is compact as a map from D into D. Schauder's Theorem [4] thus applies to T, and there exists a fixed point $f_0\in D$ which satisfies the TPBVP (17). Integrating the differential equation of (17) from zero to x yields

$$f_0'''+Q[f_0f_0''-(f_0')^2]=\beta,$$

where $\beta=f_0'''(0)-Q[f_0'(0)]^2=f_0'''(1)-Q[f_0'(1)]^2$. Since this solution lies in D, $0<f_0'(0)\leqq\frac{1}{2}$. We also proved that $f_0'''(1)=f_0^{*'''}(1)\geqq 1$. Thus $1>\beta>1-\frac{1}{4}Q$. This completes the proof of Theorem 3.

Acknowledgments. We most sincerely thank Professor S. P. Hastings for his interest and suggestions and Professor W. N. Gill for leading us to these problems.

REFERENCES

[1] S. P. HASTINGS, *An existence theorem for a problem from boundary layer theory*, Arch. Rational Mech. Anal., 33 (1969), pp. 103–109.

[2] W. N. GILL, N. D. KAZARINOFF, AND J. D. VERHOEVEN, *Convective diffusion in zone refining of low Prandtl number liquid metals and semiconductors*, in Integrated Circuits: Chemical and Physical Processing, P. Stroeve, ed., Amer. Chem. Soc. Symposium Series, No. 290, 1985, pp. 47–69.

[3] W. N. GILL, N. D. KAZARINOFF, C. C. HSU, M. A. NOACK, AND J. D. VERHOEVEN, *Thermocapillary-driven convection in supported and floating-zone driven convection*, Adv. Space Research, 4 (1984), pp. 15–22.

[4] J. SMOLLER, *Shock Waves and Reaction-Diffusion Equations*, Grundlehren der Math. Wissenschaften 258, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

009

EQUATIONS SURQUADRATIQUES ET DISPARITION DES SAUTS*

FRANCINE DIENER†

Abstract. The following question is answered: Under which conditions on f can the solutions of $\varepsilon x'' = f(t, x, x')$, ε infinitesimal, have jumps, and under which conditions will they never have jumps? To do this a nonstandard approach is used called "la methode du plan d'observabilité," which was introduced in [SIAM J. Math. Anal., 17 (1986), pp. 533-559]. The results are applied to explain the vanishing of a limit cycle and also the disappearance of the solution for some boundary value problems.

Key words. saut, surquadratique, plan d'observabilité, couche limite

AMS(MOS) subject classifications. 34A34, 34E15, 03405

Nous nous interesserons ci-dessous aux solutions $x(t)$ d'une équation différentielle du second ordre singulièrement perturbée

$$\varepsilon x'' = f(t, x, x') \tag{1}$$

où $\varepsilon > 0$ est un nombre fixé infinitésimal et f une fonction interne définie dans $\mathbb{R}^3$ est suffisamment régulière pour que l'équation différentielle possède la propriété d'existence et d'unicité des solutions.

On sait que les solutions de cette équation adoptent alternativement l'un ou l'autre des deux comportements suivants:

—D'une part un comportement qu'on peut qualifier de *lent*, correspondant à des phases pendant lesquelles la vitesse reste limitée et donc au cours desquelles la solution $x(t)$ a une ombre continue $x_0(t)$. Dans ce cas, en tout point t où elle est dérivable, l'ombre satisfait l'équation réduite $f(t, x, x') = 0$, obtenue en remplaçant ε par 0 dans l'équation.

—D'autre part un comportement *quasi-discontinu* correspondant à des phases pendant lesquelles, la vitesse atteignant des valeurs non limitées, la solution "saute" brutalement d'une phase lente à une autre phase lente.

Dans un article récent [3], nous avons montré de quelle façon on peut étudier, pour la plupart des équations (1), les sauts des solutions, leur origine, leur extrémité, leur épaisseur. Dans le présent article, nous nous proposons de préciser quelles conditions doivent être vérifiées par f pour rendre possible la présence de sauts dans les solutions et de quelle façon, lorsque ces conditions ne sont pas remplies, ces sauts peuvent être amenés à disparaître totalement, les solutions ne pouvant plus avoir alors qu'un comportement lent.

Le problème de la disparition des sauts a été rencontré pour la première fois probablement en 1952 par Coddington et Levinson [1] dans l'étude du problème aux limites

$$\varepsilon x'' = -x' - x'^3,$$

$$x(0) = a, \qquad x(1) = b.$$

Ces auteurs observèrent que pour $a \neq b$ ce problème n'a pas de solution, si ε est choisi infinitésimal. Il est facile de voir qu'aucune solution gardant un comportement lent sur tout l'intervalle de temps $[0, 1]$, ne peut convenir (car pour une telle solution on

* Received by the editors October 22, 1986; accepted for publication August 24, 1987.
† Université de Paris X, UFR SEGMI, 200 Avenue de la République, 92001 Nanterre Cedex, France.

a $-x'(t)-x'^3(t)\simeq 0$, et donc $x'(t)\simeq 0$, c'est-à-dire $x(t)$ presque constante). L'absence de solution à ce problème aux limites était donc un signe de rareté des sauts pour les solutions de cette équation, ce qui a pu paraître étrange compte tenu des seuls cas bien connus à l'époque, à savoir les cas où f est indépendante de x' ou linéaire en x'.

Cette énigme trouva son explication quelques années plus tard (1960) dans un article de Višik et Liusternik [6] où il est montré que, lorsque f se comporte, en tant que fonction de x', comme une puissance x'^s, pour x' tendant vers l'infini, il est nécessaire que s soit inférieur ou égal à 2 pour que les solutions puissent présenter des sauts.

Nous nous proposons de retrouver ce résultat, et de l'étendre à une classe très générale d'équations (1) (où f ne se comporte pas nécessairement comme une puissance de x'), et surtout d'en donner une interprétation géométrique par l'intermédiaire de la méthode du plan d'observabilité [3]. Nous verrons également le parti qu'on peut tirer d'une bonne connaissance du processus de disparition des sauts dans l'étude de problèmes à un ou plusieurs paramètres.

1. Quelques définitions. Voici quelques définitions, introduites pour la plupart dans [3], qui nous serons utiles pour la suite:

DÉFINITION. Soit $v_0\geqq 0$ limité, et soit $F:[v_0, +\infty[\to\mathbb{R}^+$ une fonction interne, de classe S^0, continuement dérivable et nulle part infinitésimale. On dit que F est le *type de croissance de* f pour les v positifs s'il existe deux fonctions internes $a(t, x)$ et $r(t, x, v)$ continues et de classe S^0 telles que, pour tout t et x limités, on ait

$$f(t, x, v)=a(t, x)F(v)+r(t, x, v) \quad \text{pour tout } v\geqq v_0 \quad \text{et}$$

$$r(t, x, v)/F(v)\simeq 0 \quad \text{pour tout } v \text{ non limité.}$$

La fonction $a(t, x)$ est la *mantisse* de f pour les v positifs.

DÉFINITION. Soit $f(t, x, v)$ une fonction ayant le type de croissance $F(v)$ pour $v\geqq v_0$. On dit que f est *surquadratique* pour les v positifs si on a

$$\int_{v_0}^{+\infty} v\,dv/F(v)<+\infty.$$

Remarque. On définirait de la même façon les type de croissance et mantisse pour les v négatifs ainsi que le fait d'être surquadratique pour les v négatifs.

Exemples. Les équations (1) de la forme $\varepsilon x''=a(t, x)x'+b(t, x)$ (équations quasilinéaires) correspondent à une fonction f ayant pour type de croissance (positif et négatif) $F(v)=v$ (non surquadratique).

Les équations de la forme $\varepsilon x''=a(t, x)x'^2+b(t, x)x'+c(t, x)$ correspondent à une fonction f ayant pour type de croissance (positif et négatif) $F(v)=v^2$ (non surquadratique).

Cependant une fonction f ayant pour type de croissance $F(v)=v^2\log v$ n'est pas non plus surquadratique.

Par contre l'équation $\varepsilon x''=-x'-x'^3$ dont nous avons parlé cidessus, correspond à une fonction surquadratique puisque son type de croissance est $F(v)=v^3$.

Remarque. Dans la suite nous supposerons toujours que l'équation différentielle (1) est définie par une fonction f ayant un type de croissance pour les v positifs ainsi que pour les v négatifs (ces deux types de croissance seront presque toujours identiques dans les exemples). On se limitera également, sauf mention contraite, au cas des vitesses positives, celui des vitesses négatives s'en déduisant facilement.

DÉFINITION. Un instant standard t_0 sera dit *singulier* pour l'équation $\varepsilon x''=f(t, x, x')$ s'il existe un intervalle standard $[x_1, x_2]$, tel que pour tout $x\in[x_1, x_2]$, $a(t_0, x)\simeq 0$.

DÉFINITION. On dit qu'une solution $x(t): I \to \mathbb{R}$ (qu'on supposera toujours limitée) *présente un saut sur l'intervalle* $[t_1, t_2] \subset I$ si, sur cet intervalle, la vitesse $x'(t)$ est non limitée et si $x(t_1) \neq x(t_2)$. Si t_0 est la partie standard commune de t_1 et t_2, on dit aussi que $x(t)$ *présente un saut à l'instant* t_0. Enfin un saut à l'instant t_0 sera dit *singulier* si t_0 est un instant singulier pour l'équation et si l'intervalle standard $[x_1, x_2]$ sur lequel $a(t_0, x) \simeq 0$ rencontre l'intervalle $[x(t_1), x(t_2)]$. Sinon, le saut est dit *régulier.*

2. La géométrie des sauts: les plans d'observabilité. Pour étudier les sauts des solutions d'une équation (1), on pense généralement à un changement d'échelle de temps. En effet un saut étant un passage presque instantané ($t_1 \simeq t_2$) d'une position ($x(t_1)$) à une autre position ($x(t_2)$), il est naturel d'imaginer qu'un changement d'échelle de temps doit permettre, en "étalant" le saut dans le temps, d'en faciliter l'étude. C'est l'idée de départ de la plupart des études classiques [4], [5]. Malheureusement, si les changements de temps auxquels on pense aussitôt, tels que $T = t/\varepsilon$, ou $T = t/\sqrt{\varepsilon}$, se révèlent efficaces pour certains types d'équations (1), comme par exemple lorsque f est indépendante de x' ou linéaire en x', ils ne conviennent pas pour l'étude des sauts pour la plupart des autres équations.

La méthode du plan d'observabilité [2], [3], envisage l'étude des sauts d'un point de vue différent. Il consiste à remplacer l'équation (1) par le champ de vecteurs lent-rapide de $\mathbb{R}^3$

$$\mathcal{V}\begin{cases} t' = 1, \\ x' = v, \\ \varepsilon v' = f(t, x, v), \end{cases}$$

et à étudier les portions de trajectoires de ce champ qui correspondent aux sauts de l'équation initiale. Avec ce point de vue géométrique (Fig. 1) il apparaît aussitôt que pour étudier les sauts, qui n'ont pas d'ombre à l'échelle initiale puisqu'au cours du saut v est non limité, il est naturel de faire un changement d'échelle, de vitesse cette fois. On pose $v = h(V/\varepsilon)$ et on cherche à déterminer h en fonction de f de telle sorte

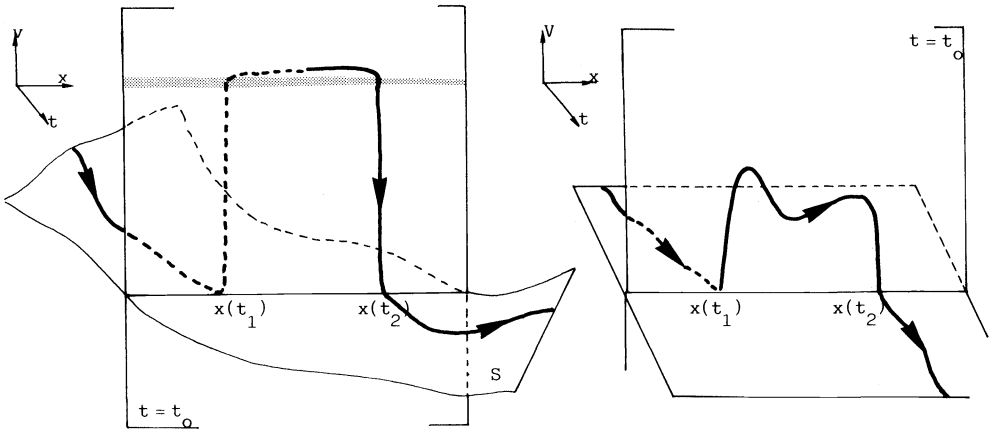


FIG. 1. *Le champ de vecteur* $\mathcal{V}$ *associé à l'équation* (1) *est presque vertical hors du halo de la surface lente S et aussi longtemps que l'ordonnée v reste limitée. On a représenté une trajectoire de* $\mathcal{V}$*: elle présente une portion lente contenue dans le halo de S, deux portions verticales et un saut contenus dans le halo du plan* $t = t_0$ *au cours desquels la trajectoire passe presque instantanément de la valeur* $x(t_1)$ *à la valeur* $x(t_2)$ *avec une vitesse non limitée.*

Après changement d'échelle de vitesse $v = h(V/\varepsilon)$*, le saut reste d'ordonnée limitée. Son ombre est contenue dans le plan d'observabilité* $t = t_0$ *et elle satisfait l'équation* (*).

que le champ obtenu à la nouvelle échelle soit, après division de ses composantes par la quantité $h(V/\varepsilon)$, un champ presque standard intégrable (voir ci-dessous la démonstration du théorème). On constate que si f a pour $v \geqq v_0$ le type de croissance $F(v)$, il suffit de prendre pour h la solution de l'équation différentielle $h' = F(h)/h$ telle que $h(0) = v_0$. A la nouvelle échelle, un saut à l'instant t_0 a pour ombre la courbe, contenue dans le plan $t = t_0$, d'équation

$$(*) \qquad V(x) = V(x_0) + \int_{x_0}^{x} a(t_0, \xi)\, d\xi$$

où $a(t, x)$ est la mantisse de f. Les plans $t = t_0$ dans lesquels viennent se "ranger" les ombres des sauts à l'échelle (x, V) sont les plans d'observabilité des sauts [2].

Notons encore que, pour les sauts singuliers, le plan d'observabilité n'est pas l'échelle adhéquate, puisque, $a(t, x)$ s'annulant pour $t = t_0$, les sauts ont tous pour ombre à cette échelle des droites horizontales $V(x) = V_0$ (Fig. 2). Donc si une trajectoire issue d'un point de coordonnées limitées à l'échelle initiale présente un saut singulier, $V(x)$ reste infinitésimal tout au long du saut. L'examen des trajectoires des plans d'observabilité ne permet donc pas, dans ce cas, de décrire les sauts (origine et extrémité, par exemple) ni même de prouver leur existence.

3. Disparition des sauts.

THÉORÈME. *Soit $f(t, x, v)$ une fonction ayant pour type de croissance pour $v > 0$ la fonction $F(v)$. Supposons $F(v)$ surquadratique et posons $V_0 = \varepsilon \int_0^{+\infty} v\, dv/F(v)$. Alors tout saut $x(t)$ de l'équation différentielle $\varepsilon x'' = f(t, x, x')$ doit satisfaire à la fois*:

(a) $V(x(t)) = V(x_0) + \int_{x_0}^{x(t)} a(t_0, \xi)\, d\xi$,

(b) $V(x(t)) \lesssim V_0$,

où $V(x(t))$ est défini par $x'(t) = h(V(x(t))/\varepsilon)$ avec $h' = F(h)/h$ et $h_0 = v_0$.

COROLLAIRE. *Aucune solution d'une équation différentielle* (1) *pour laquelle f est standard et de type de croissance surquadratique ne peut présenter de saut régulier.*

Preuve du théorème. La preuve s'appuie sur la remarque suivante: l'équation différentielle $h' = F(h)/h$, permettant de déterminer un changement d'échelle $h(w)$ convenable, a des solutions globales, c'est-à-dire définies pour toutes valeurs de la variables w, si et seulement si F n'est pas surquadratique. Au contraire, lorsque F est surquadratique, la solution $h(w)$ de cette équation telle que $h(0) = v_0$ est définie sur l'intervalle $[0, w_0[$ où $w_0 = \int_0^{+\infty} v\, dv/F(v)$ et elle tend vers $+\infty$ quand w tend vers $w_0 = V_0/\varepsilon$.

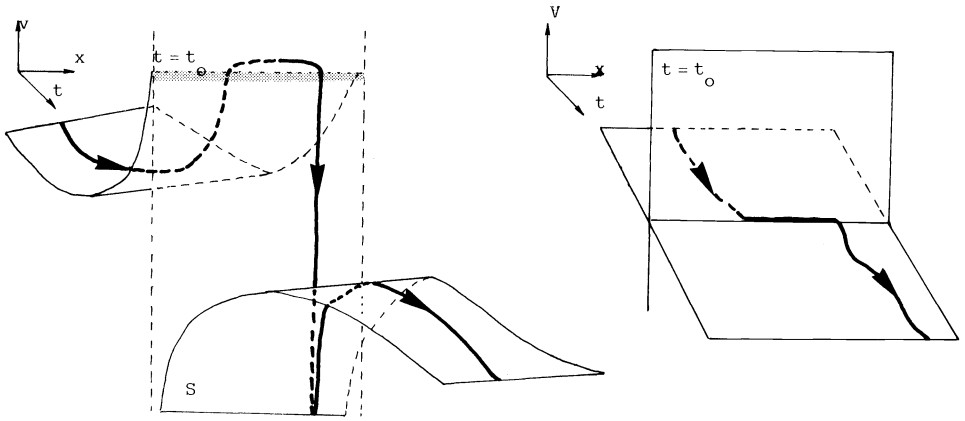


FIG. 2. *Dans le cas d'un saut singulier, la surface S est asymptote au plan $t = t_0$ entre $x(t_1)$ et $x(t_2)$. A l'échelle du plan d'observabilité, le saut a une ordonnée infinitésimale entre ces deux abscisses.*

Effectuons le changement d'échelle $v = h(V/\varepsilon)$. Le champ v se transforme en

$$\begin{cases} t' = 1, \\ x' = h(V/\varepsilon), \\ V' = f(t, x, h(V/\varepsilon))/h'(V/\varepsilon) \end{cases}$$

qui a mêmes trajectoires, lorsque $v(=h(V/\varepsilon))$ est non limité (et donc non nul) que le champ:

$$\tilde{\mathcal{V}} \begin{cases} t' = 1/h(V/\varepsilon), \\ x' = 1, \\ V' = f(t, x, h(V/\varepsilon))/h(V/\varepsilon)h'(V/\varepsilon) \end{cases}$$

qui est presque égal, lorsque v est non limité, au champ standard de composantes $(0, 1, a(t, x))$, en vertu des hypothèses sur f et h. On en déduit, comme dans le cas où f n'est pas surquadratique [3], que les ombres des trajectoires satisfont l'équation (a). De plus, comme h n'est, dans le cas surquadratique, définie que sur l'intervalle $[0, V_0/\varepsilon[$, seules les portions de trajectoires de $\tilde{\mathcal{V}}$ pour lesquelles V reste inférieur à V_0 sont les images de trajectoires du champ initial (Fig. 3). D'où l'inégalité (b).

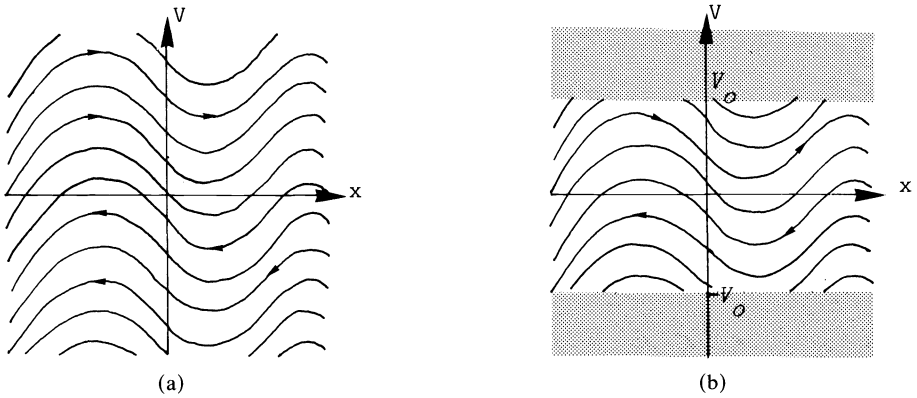


FIG. 3. *Ombres des sauts dans le plan d'observabilité.* (a) *Cas où f n'est pas surquadratique.* (b) *Cas où f est surquadratique* ($V_0 = \varepsilon \int_0^\infty v\, dv/F(v)$).

Preuve du corollaire. C'est une conséquence immédiate du théorème. En effet dans le cas où f est standard, $w_0 = \int_0^{+\infty} vdv/F(v)$ l'est également et donc $V_0 = \varepsilon w_0$ est infinitésimal. De l'inégalité (b) il résulte donc que, à la nouvelle échelle, on a $V(x(t)) \simeq 0$. Or si $x(t)$ est un saut régulier à l'instant t_0, $a(t_0, x)$ reste appréciable (i.e., non infinitésimal), ce qui est impossible compte tenu de l'égalité (a).

Commentaires. (1) Le corollaire ci-dessus pourrait être reformulé en termes plus géométriques de la façon suivante: lorsque f est standard et à type de croissance surquadratique, et si $a(t, x)$ ne s'annule identiquement pour aucune valeur de t, les trajectoires de $\mathcal{V}$ restent, à l'échelle initiale, équivalentes à leur ombre, pourvu que $a(t, x)$ ne s'annulle pas identiquement pour certaines valeurs de t. Ceci est faux si les trajectoires peuvent présenter des sauts car dans ce cas, même si leurs ombres sont verticales à l'échelle initiale, les trajectoires ne sont pas quasi verticales lorsque v est non limité. Dans le cas surquadratique cependant, le corollaire précédent montre que les trajectoires restent quasi verticales même pour des v non limités.

(2) Les résultats ci-dessus ne disent rien sur l'existence éventuelle de sauts singuliers lorsque f est surquadratique. En fait, de tels sauts peuvent fort bien exister, comme l'a montré Howes par exemple [4] aussi bien dans le cas surquadratique que dans le cas contraire. Leur étude complète reste à faire.

(3) Le corollaire ci-dessus est une nouvelle formulation, dans un cas un peu plus général, des résultats de Višik et Liusternik. Mais, comme ceux-ci, il donne de la disparition des sauts, lorsque le type de croissance devient surquadratique, l'impression d'un phénomène discontinu: si le type de croissance n'est pas surquadratique, toutes les trajectoires standard du champ de composantes $(0, 1, a(t, x))$ sont ombres de sauts, s'il le devient, aucune n'est plus l'ombre d'un saut de l'équation initiale. En réalité, la disparition des sauts lorsque f devient surquadratique, est progressive. Mais pour s'en rendre compte, il convient de prendre en considération également certaines équations non standard intermédiaires: plus précisément celles qui sont surquadratiques mais pour lesquelles $w_0 = \int_0^{+\infty} v/F(v)\, dv$ est non limité. L'étude de familles à un paramètre, proposée au paragraphe suivant, illustrera ce mécanisme de disparition progressive.

4. Exemples de bifurcations liés à la disparition des sauts. (a) Considérons la famille d'équations autonomes à un paramètre suivante

$$\varepsilon x'' + (x^2 - 1)x'^{[s]} + x = 0$$

où $x'^{[s]}$ désigne la fonction impaire de x' égale à x'^s quand $x' \geqq 0$ (étudiée dans [2], [3]). Les trajectoires du champ $\mathcal{V}$ associé se projettent, puisque l'équation ne dépend pas du temps, sur celles du champ de vecteurs du plan

$$x' = v,$$
$$\varepsilon v' = (1 - x^2)v^{[s]} - x.$$

Le type de croissance $F_s(x) = v^s$ est surquadratique si et seulement si $s > 2$. Lorsque $s = 1$ (équation de van der Pol) il existe un unique cycle limite vers lequel tendent toutes les trajectoires, à l'exception du point stationnaire $x = v = 0$. Au contraire lorsque s est standard et strictement supérieur à 2, les trajectoires restent, en vertu du commentaire (1) ci-dessus, équivalentes à leurs ombres à l'échelle initiale: il en résulte qu'il ne peut y avoir de cycle limite dans ce cas. Comment et pour quelle valeur du paramètre le cycle disparait-il?

Dans le plan d'observabilité (x, V), où V est défini par $v = h(V/\varepsilon)$ comme précédemment, les sauts, s'ils existent, ont nécessairement pour ombre les courbes d'équation

$$V(x) = V(x_0) + \int_{x_0}^{x} (1 - \xi^2)\, d\xi$$

ou bien encore

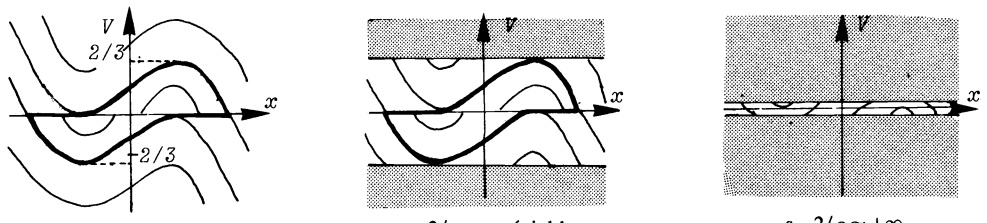
$$V(x) = x - x^3/3 + K, \quad K \text{ constante.}$$

D'autre part, dès que s est supérieur à 2, il convient de ne considérer que les portions de ces courbes dont l'ordonnée ne dépasse pas, en valeur absolue,

$$V_0 = \varepsilon w_0 = \varepsilon \int_{v_0}^{+\infty} v/v^s\, dv = \varepsilon v_0^{2-s}/(s-2).$$

On peut choisir $v_0 = 1$. Le cycle limite, qui doit nécessairement, s'il existe, atteindre la vitesse $V = 2/3$ (et $V = -2/3$) comme l'indique la Fig. 4, n'existe qu'à la condition que $\varepsilon/(s-2) \gg 2/3$; en d'autres termes, il disparaît, en "éclatant" à l'infini, pour une valeur s_0 telle que

$$s_0 \simeq 2 + 3\varepsilon/4.$$



$s-2/\varepsilon \simeq 0$ $s-2/\varepsilon$ appréciable $s-2/\varepsilon \simeq +\infty$

FIG. 4

Mais auparavant, alors que le cycle était un attracteur global (excepté pour le point stationnaire) pour $s \leqq 2$, il existe, dès que $s > 2$ certaines trajectoires qui ne tendent plus vers le cycle mais qui tendent vers l'infini. Le nombre de ces trajectoires augmente progressivement avec s jusqu'à la disparition du cycle.

(b) Le second exemple concerne non plus une bifurcation de portrait de phase mais la disparition de la solution d'un problème aux limites lorsque le type de croissance de l'équation devient surquadratique.

Considérons le problème aux limites suivant:

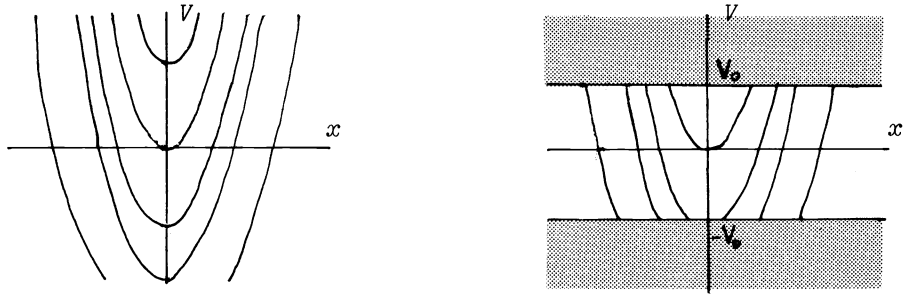
$$\varepsilon x'' = x x'^{[s]},$$
$$x(-1) = a,$$
$$x(1) = b.$$

Comme pour l'exemple historique évoqué en introduction, il est facile de voir qu'aucune solution de l'équation, lente sur tout l'intervalle $[-1, 1]$, ne peut satisfaire à la fois les deux conditions aux limites, sauf éventuellement si $a \simeq b$, car les solutions lentes sont presque constantes (x' nul). Donc lorsque s est standard et strictement supérieur à 2, ce problème ne peut avoir de solutions si $a \neq b$. Par ailleurs des études classiques [5], [4], ou l'examen des sauts dans leur plan d'observabilité, montrent qu'il existe une solution lorsque $s = 1$ et même plus généralement lorsque $0 < s \leqq 2$. Que peut-on dire du problème posé lorsque s est supérieur et équivalent à 2?

Dans le plan d'observabilité (x, V), où $V = h(v/\varepsilon)$ comme précédemment, les sauts, s'ils existent, ont nécessairement pour ombres les courbes d'équation

$$V(x) = V(x_0) + \int_{x_0}^{x} \xi \, d\xi$$

ou bien encore $V(x) = x^2/2 + K$, K constante (Fig. 5). D'autre part, dès que s est



$s \leqq 2$ $s > 2, 0 \ll (s-2)/\varepsilon \ll +\infty$.

FIG. 5

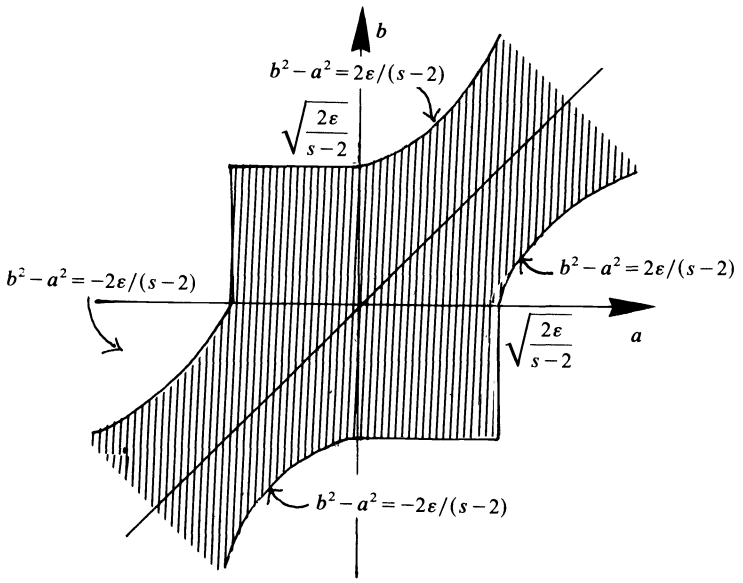


FIG. 6

supérieur à 2, il convient de ne considérer que les portions de courbes dont l'ordonnée ne dépasse pas, en valeur absolue,

$$V_0=\varepsilon w_0=\varepsilon\int_0^\infty v/v^s\,dv=\varepsilon v_0^{2-s}/(s-2)=\varepsilon/(s-2)\quad\text{pour } v_0=1.$$

Il est alors facile de voir à quelles conditions sur a et b il est possible ou non de joindre, en longeant une parabole d'équation $V=x^2/2+K$, un point d'abscisse a à un point d'abscisse b sans dépasser l'ordonnée $\varepsilon/(s-2)$. Plus précisément, pour s fixé, strictement supérieur à 2, le problème aux limites posé possède une solution si et seulement si a et b appartiennent à la région hachurée indiquée sur la Fig. 6. Cette région tend à recouvrir le plan tout entier lorsque s tend vers 2 et tend à se réduire à la diagonale $a=b$ lorsque s croît au-delà de 2.

REFERENCES

[1] E. A. CODDINGTON AND N. LEVINSON, *A boundary value problem for a nonlinear differential equation with a small parameter*, Proc. Amer. Math. Soc., 3, 1952, pp. 73-81.

[2] F. DIENER, *Méthode du plan d'observabilité*, Thèse, Strasbourg, 1981.

[3] ———, *Sauts des solutions des équations* $\varepsilon x''=f(t,x,x')$, SIAM J. Math. Anal., 17 (1986), pp. 533-559.

[4] F. A. HOWES, *Boundary-interior layer interactions in nonlinear singular perturbation theory*, Mem. Amer. Math. Soc., 203 (1978), pp. 1-108.

[5] R. E. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[6] H. I. VIŠIK AND L. A. LIUSTERNIK, *Initial jump for nonlinear differential equations containing a small parameter*, Soviet Math. Dokl., 1 (1960), pp. 719-752.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

010

SINGULAR SELF-ADJOINT STURM–LIOUVILLE PROBLEMS. II: INTERIOR SINGULAR POINTS*

ALLAN M. KRALL† AND ANTON ZETTL‡

Abstract. The second-order Sturm–Liouville operator

$$ly = [-(py')' + qy]/w$$

is considered over a region (a, b) on the real line, $-\infty \leqq a < b \leqq \infty$, on which the operator may have a finite number of singular points. By considering l over various subintervals on which singularities occur only at the ends, restrictions of the maximal operator generated by l in $L^2_w(a, b)$ may be found which are self-adjoint. In addition to direct sums of self-adjoint operators defined on the separate subintervals, there are other self-adjoint restrictions of the maximal operator which involve linking the various intervals together in interface-like style.

Key words. singular, Sturm–Liouville, boundary condition, operator adjoint

AMS(MOS) subject classifications. 34A30, 34B05, 34B10, 34B20, 34B25

1. Introduction. This article is an extension of the work of Everitt and Zettl [2], which dealt with the problem of finding self-adjoint operators of the form

$$ly = [-(py')' + qy]/w$$

with one interior singular point, or possibly over two disjoint intervals, using singular Naimark boundary forms [8]. We use the *equivalent* concrete boundary representation discussed in [5] and consider finitely many singular points, or perhaps finitely many disjoint intervals.

The extension to many singular points or many disjoint intervals is done with relative ease because we use the explicit Fulton-type boundary forms exhibited in [3], [5], [6] for singular ends. By using these concrete forms, not only are direct sum self-adjoint operators easily exhibited, but also self-adjoint operators whose boundaries are linked together are explicitly described. We also bypass the abstract and rather difficult to use Naimark boundary forms found in [8].

We assume that the terminology of limit-point and limit-circle ends is familiar to the reader. Classic descriptions may be found in [1], [4], [9], as well as many other books on differential equations. Essentially limit-point means that the differential equation

$$-(py')' + qy = \lambda wy, \qquad \operatorname{Im} \lambda \neq 0$$

has only one independent solution that is square integrable in any local region containing the singular point. Limit-circle implies that all solutions are locally square integrable for *all* λ near the singular point.

Regular endpoints may be thought of as benign limit-circle points.

* Received by the editors May 25, 1987; accepted for publication (in revised form) October 21, 1987. This work was partially supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract W-31-109-Eng-38.

† Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, and Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

‡ Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, and Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

We can without loss of generality assume that the interval (a, b), $-\infty \leqq a < b \leqq \infty$, in question is decomposed into four sets of subintervals:

(1) $\{I_j\}_{j=1}^m$. Considered on I_j, l is limit-point at both ends.

(2) $\{J_j\}_{j=1}^n$. Considered on J_j, l is limit-point at the left end, limit-circle at the right end.

(3) $\{K_j\}_{j=1}^p$. Considered on K_j, l is limit-circle at the left end, limit-point at the right end.

(4) $\{L_j\}_{j=1}^q$. Considered on L_j, l is limit-circle at both ends.

DEFINITION 1.1. We denote by D_M the collection of those elements y satisfying the following:

(1) $y \in L_w^2(I_j)$, $j = 1, \cdots, m$, $y \in L_w^2(J_j)$, $j = 1, \cdots, n$, $y \in L_w^2(K_j)$, $j = 1, \cdots, p$, $y \in L_w^2(L_j)$, $j = 1, \cdots, q$.

(2) y is differentiable almost everywhere in each I_j, J_j, K_j, L_j. (py') is locally absolutely continuous in each I_j, J_j, K_j, L_j.

(3) ly exists in each I_j, J_j, K_j, L_j by 2, and $ly \in L_w^2(I_j)$, $j = 1, \cdots, m$, $ly \in L_w^2(J_j)$, $j = 1, \cdots, n$, $ly \in L_w^2(K_j)$, $j = 1, \cdots, p$, $ly \in L_w^2(L_j)$, $j = 1, \cdots, q$.

DEFINITION 1.2. We define the operator L_M by setting $L_M y = ly$ for all $y \in D_M$.

The underlying Hilbert space is, of course,

$$H = \sum_{j=1}^{m} L_w^2(I_j) \oplus \sum_{j=1}^{n} L_w^2(J_j) \oplus \sum_{j=1}^{p} L_w^2(K_j) \oplus \sum_{j=1}^{q} L_w^2(L_j).$$

2. Green's formulas. In order to properly look for restrictions of L_M, Green's formula for each of the regions I_j, J_j, K_j, L_j must be developed. It is by using the sum of these that the restrictions *through boundary conditions* can be developed.

Let us consider I_j, and let (α, β) be a subinterval of I_j with neither α nor β an end of I_j. It is an easy computation to show that if $y, z \in D_M$, then

$$\int_\alpha^\beta [\bar{z}(L_M y) - (\overline{L_M z})y]w\,dx = p[y\bar{z}' - y'\bar{z}]|_\alpha^\beta.$$

Likewise it is well known that as x approaches a limit-point end, $p[y\bar{z}' - y'\bar{z}]$ approaches zero. In this case, therefore,

$$\int_{I_j} [\bar{z}(L_M y) - (\overline{L_M z})y]w\,dx = 0, \qquad j = 1, \cdots, m.$$

Now replace I_j by J_j. If $J_j = (\alpha_j, \beta_j)$, then as α approaches α_j, $p[y\bar{z}' - y'\bar{z}]$ approaches zero; but as β approaches β_j, it does not necessarily approach zero. A closer look is required. Note that $p[y\bar{z}' - y'\bar{z}]$ can be written as

$$(\bar{z}, p\bar{z})\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} y \\ py' \end{pmatrix}.$$

Let θ, ϕ be solutions of $ly = 0$ satisfying $p(\theta\phi' - \theta'\phi) = 1$. Then

$$-\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \theta & \phi \\ p\theta' & p\phi' \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \theta & p\theta' \\ \phi & p\phi' \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

If this is inserted in the middle of the preceding product, the result is

$$(\overline{W(z, \theta)}, \overline{W(z, \phi)})\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} W(y, \theta) \\ W(y, \phi) \end{pmatrix}$$

where $W(f, g) = p(fg' - f'g)$.

Furthermore, since both θ and ϕ are square integrable near β_j, the terms W all have finite limits as β approaches β_j. (Use Green's formula, or see [7].) Hence Green's formula over J_j becomes

$$\int_{J_j} [\bar{z}(L_M y) - (\overline{L_M z})y]w\,dx = (\overline{Q_j(z,\theta)}, \overline{Q_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}Q_j(y,\theta)\\ Q_j(y,\phi)\end{pmatrix},$$

where Q replaces W to indicate the limit has been taken as β approaches β_j.

If the interval is K_j, rather than J_j, then it is the lower limit as α approaches α_j that remains. With the limit-circle case holding at the lower end, therefore,

$$\int_{K_j} [\bar{z}(L_M y) - (\overline{L_M z})y]w\,dx = -(\overline{R_j(z,\theta)}, \overline{R_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}R_j(y,\theta)\\ R_j(y,\phi)\end{pmatrix}$$

where R replaces W to indicate the limit has been taken as α approaches α_j.

Finally if the interval is L_j, limiting terms at both ends remain. If S indicates a limit at β, and T a limit at α, then

$$\int_{L_j} [\bar{z}(L_M y) - (\overline{L_M z})y]w\,dx = (\overline{S_j(z,\theta)}, \overline{S_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}S_j(y,\theta)\\ S_j(y,\phi)\end{pmatrix}$$
$$-(\overline{T_j(z,\theta)}, \overline{T_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}T_j(y,\theta)\\ T_j(y,\phi)\end{pmatrix}.$$

Green's formula over all of (a, b) is the sum of these. If we let $\langle\cdot,\cdot\rangle$ denote the inner product over H,

$$\langle f, g\rangle = \sum_{j=1}^{m}\int_{I_j} \bar{g}fw\,dx + \sum_{j=1}^{n}\int_{J_j} \bar{g}fw\,dx + \sum_{j=1}^{p}\int_{K_j} \bar{g}fw\,dx + \sum_{j=1}^{q}\int_{L_j} \bar{g}fw\,dx,$$

then summing the previous expressions, we get the following theorem.

THEOREM 2.1. *Let $y, z \in D_M$. Then,*

$$\langle L_M y, z\rangle - \langle y, L_M z\rangle = \sum_{j=1}^{n}(\overline{Q_j(z,\theta)}, \overline{Q_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}Q_j(y,\theta)\\ Q_j(y,\phi)\end{pmatrix}$$
$$-\sum_{j=1}^{m}(\overline{R_j(z,\theta)}, \overline{R_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}Q_j(y,\theta)\\ Q_j(y,\phi)\end{pmatrix}$$
$$+\sum_{j=1}^{q}(\overline{S_j(z,\theta)}, \overline{S_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}S_j(y,\theta)\\ S_j(y,\phi)\end{pmatrix}$$
$$-\sum_{j=1}^{q}(\overline{T_j(z,\theta)}, \overline{T_j(z,\phi)})\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\begin{pmatrix}T_j(y,\theta)\\ T_j(y,\phi)\end{pmatrix}.$$

This is Green's formula over all of (a, b).

3. General boundary conditions. The sums involved in Green's formula may be more efficiently handled by the use of additional matrix notation. Let $B(y)$, $B(z)$, and $\mathbf{J}$ be defined as follows:

$$B(y) = (Q_1(y,\theta)\cdots Q_n(y,\theta), R_1(y,\theta)\cdots R_p(y,\phi), S_1(y,\theta)\cdots S_q(y,\phi),$$
$$T_1(y,\theta)\cdots T_q(y,\phi))^T,$$
$$B(z) = (Q_1(z,\theta)\cdots Q_n(z,\phi), R_1(z,\theta)\cdots R_p(z,\phi), S_1(z,\theta)\cdots S_q(z,\phi),$$
$$T_1(z,\theta)\cdots T_q(z,\phi))^T.$$

Here θ and ϕ terms alternate, giving first Q, then R, then S, then T terms.

$$\mathbf{J}=\begin{pmatrix}\mathbf{J}_1 & 0 & 0 & 0\\ 0 & \mathbf{J}_2 & 0 & 0\\ 0 & 0 & \mathbf{J}_3 & 0\\ 0 & 0 & 0 & \mathbf{J}_4\end{pmatrix},$$

where

$$\mathbf{J}_1=\begin{pmatrix}J & & \\ & \ddots & \\ & & J\end{pmatrix}$$

consists of n blocks of

$$J=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}.$$

$$\mathbf{J}_2=\begin{pmatrix}-J & & \\ & \ddots & \\ & & -J\end{pmatrix}$$

consists of p blocks of $-J$. $\mathbf{J}_3$ is like $\mathbf{J}_1$, but consists of q blocks. $\mathbf{J}_4$ is like $\mathbf{J}_2$, but consists of q blocks.

THEOREM 3.1. *Green's formula for $y, z \in D_M$ is*

$$\langle L_M y, z\rangle - \langle y, L_M z\rangle = B(z)^*\mathbf{J}B(y).$$

General boundary conditions involve linear combinations of terms Q_j, R_j, S_j, T_j, or, more concisely, combinations of the entries in $B(y)$. These are introduced by matrix multiplication.

Let M be an $r\times(2n+2p+4q)$ matrix, rank $M=r$. Let N be a $(2n+2p+4q-r)\times(2n+2p+4q)$ matrix, rank $N=2n+2p+4q-r$. Let $\binom{M}{N}$ be nonsingular.

Likewise let P be an $r\times(2n+2p+4q)$ matrix, rank $P=r$. Let Q be a $(2n+2p+4q-r)\times(2n+2p+4q)$ matrix, rank $Q=2n+2p+4q-r$. Assume also that

$$(P^*, Q^*)\begin{pmatrix}M\\ N\end{pmatrix}=\mathbf{J}.$$

THEOREM 3.2. *Green's formula for $y, z \in D_M$ is*

$$\langle L_M y, z\rangle - \langle y, L_M z\rangle = [PB(z)]^*[MB(y)]+[QB(z)]^*[NB(y)].$$

The proof consists of substituting for $\mathbf{J}$, and carrying out the matrix multiplication.

4. Restrictions of L_M, self-adjointness. We are now in a position to restict L_M by the imposition of boundary conditions.

DEFINITION 4.1. We denote by D the collection of those elements y satisfying the following:

(1) $y\in D_M$;

(2) $MB(y)=0$.

DEFINITION 4.2. We define the operator L by setting $Ly=ly$ for all $y\in D$.

DEFINITION 4.3. We denote by D^* the collection of those elements z satisfying the following:

(1) $z\in D_M$;

(2) $QB(z)=0$.

DEFINITION 4.4. We define the operator L^* by setting $L^*z=lz$ for all z in D^*.

We have abused notation here because traditionally L^* denotes the adjoint operator in H. We clear this up immediately.

THEOREM 4.5. *The adjoint of L in H is L^*. Likewise the adjoint of L^* in H is L.*

Proof. It is well known that the form of the adjoint of L is l (see [6]). Green's formula shows that if $MB(y)=0$, while $NB(y)$ is arbitrary, then $QB(z)=0$.

Conversely, the operator with form lz and domain D^* is clearly contained in the adjoint of L. So the adjoint is L^*.

We show that $(L^*)^*$ is L in the same way.

There are parametric forms for the boundary conditions as well. In order to characterize self-adjointness, these forms are used here.

We have

$$\begin{pmatrix} M \\ N \end{pmatrix} B(y) = \begin{pmatrix} 0 \\ \Delta \end{pmatrix},$$

where Δ is arbitrary. If this is multiplied by $-\mathbf{J}(P^*, Q^*)$, then since $\mathbf{J}^2 = -I$,

$$B(y) = -\mathbf{J}(P^*, Q^*)\begin{pmatrix} 0 \\ \Delta \end{pmatrix},$$

or

$$B(y) = -\mathbf{J}Q^*\Delta.$$

This parametric boundary condition is equivalent to $MB(y)=0$.

Likewise, if

$$B(z)^*(P^*, Q^*) = (\Gamma^*, 0)$$

where Γ is arbitrary, then postmultiplying by $-\binom{M}{N}\mathbf{J}$ yields

$$B(z) = \mathbf{J}M^*\Gamma$$

as the adjoint parametric boundary conditions, equivalent to $QB(z)=0$.

THEOREM 4.6. *L is self-adjoint if and only if $r = n+p+2q$ and $M\mathbf{J}M^* = 0$.*

Proof. If L is self-adjoint, then the number of boundary conditions for L and L^* is the same. Hence $2n+2p+4q-r=r$. Furthermore, z must satisfy the D boundary condition, so

$$MB(z) = M\mathbf{J}M^*\Gamma = 0.$$

Since Γ is arbitrary, $M\mathbf{J}M^* = 0$.

Conversely, if $r = n+p+2q$ and $M\mathbf{J}M^* = 0$, then the number of boundary constraints is the same. Further, since

$$(P^*, Q^*)\begin{pmatrix} M \\ N \end{pmatrix} = \mathbf{J},$$

we have

$$-\mathbf{J}(P^*, Q^*)\begin{pmatrix} M \\ N \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix},$$

and reversing the order,

$$\begin{pmatrix} M \\ N \end{pmatrix}(-\mathbf{J}P^*, -\mathbf{J}Q^*) = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

This implies $-MJQ^* = 0$. This further implies that there is a nonsingular matrix C such that $Q^* = M^*C^*$ or $Q = CM$. Thus $QB(y) = 0$ and $MB(y) = 0$ are equivalent boundary conditions.

In view of the connection made in [5], the following statement can be made.

THEOREM 4.7. *Let M be an $(n+p+2q)\times(2n+2p+4q)$ matrix satisfying $MJM^* = 0$. Then L, defined by Definition 4.2, is self-adjoint. Conversely, if L is a self-adjoint differential operator which is a restriction of L_M, then there exists a matrix M, with the above-mentioned properties, such that the domain of L is restricted by $MB(y) = 0$ as in Definition 4.1.*

5. Examples. Let us assume that $m = 0$, $n = 2$, $p = 0$, $q = 0$. Suppose that (a, b) consists of $(0, 2)$ with an interior singularity at $x = 1$. Suppose further that l is limit-point at 0 and 1+, but limit-circle at 1− and 2. Thus $J_1 = (0, 1)$, $J_2 = (1, 2)$, and

$$B(y) = (Q_1(y, \theta), Q_1(y, \phi), Q_2(y, \theta), Q_2(y, \phi))^T.$$

Simple separated boundary conditions are given by $MB(y) = 0$, where

$$M = \begin{pmatrix} \alpha & \beta & 0 & 0 \\ 0 & 0 & \gamma & \delta \end{pmatrix}$$

where α, β, γ, δ are real, $\alpha^2+\beta^2 \neq 0$, $\gamma^2+\delta^2 \neq 0$. This problem is equivalent to the direct sum of two self-adjoint problems, one on $(0, 1-)$, one on $(1+, 2)$, joined together.

A new problem in which the intervals are mixed together would be generated by

$$M = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 \end{pmatrix}.$$

Here, the intervals cannot be separated.

As a second example let $m = 0$, $n = 1$, $p = 0$, $q = 1$. Suppose that (a, b) consists of $(0, 2)$ with an interior singular point at $x = 1$, l being limit-point at 0, limit-circle at $1-$, at $1+$ and at 2. Thus $J_1 = (0, 1)$, $L_1 = (1, 2)$. Then $B(y)$ is given by

$$B(y) = (Q_1(y, \theta), Q_1(y, \phi), S_1(y, \theta), S_2(y, \phi), T_1(y, \theta), T_1(y, \phi))^T.$$

A general set of self-adjoint, mixed boundary conditions is given by

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 4 & 3 & 5 \\ 2 & 4 & 6 & 8 & 7 & 11 \end{pmatrix}.$$

We close with two classic examples. First consider the Legendre operator

$$ly = ((1-x^2)y')'.$$

$\pm\infty$ are limit-point, and no boundary conditions are required at those points. ± 1 from either side are limit-circle, however. Thus here, $m = 0$, $n = 1$, $p = 1$, $q = 1$, $J_1 = (-\infty, -1-)$, $K_1 = (1+, \infty)$, and $L_1 = (-1+, 1-)$.

$$H = L^2(-\infty, -1) \oplus L^2(1, \infty) \oplus L^2(-1, 1).$$

Boundary terms $B(y)$ are given by

$$B(y) = (Q_1(y, \theta), Q_1(y, \phi), R_1(y, \theta), R_1(y, \phi), S_1(y, \theta), S_1(y, \phi), T_1(y, \theta), T_1(y, \phi))^T,$$

where $\theta = 1$ in all intervals, $\phi = \frac{1}{2}\ln((x-1)/(x+1))$ on $(-\infty, -1)$ and $(1, \infty)$, but $\phi = \frac{1}{2}\ln((1+x)/(1-x))$ on $(-1, 1)$.

With $\mathbf{J} = \operatorname{diag}(J, -J, J, -J)$,

$$\langle L_M y, z\rangle - \langle y, L_M z\rangle = [PB(z)]^*[MB(y)] + [QB(z)]^*[NB(y)]$$

provided $(P^*, Q^*)\binom{M}{N} = \mathbf{J}$.

Self-adjointness occurs when $M\mathbf{J}M^* = 0$. The simplest case is that of separated conditions. Since M is 4×8, let $m_{11} = m_{23} = m_{35} = m_{47} = 1$, with $m_{ij} = 0$ otherwise. The four boundary terms produced are $Q_1(y, 1) = 0$, $R_1(y, 1) = 0$, $S_1(y, 1) = 0$, $T_1(y, 1) = 0$, which are satisfied by the Legendre polynomials. The projection onto the last component (in $L^2(-1, 1)$) generates the self-adjoint boundary value problem traditionally associated with the Legendre polynomials.

The Laguerre operator

$$ly = -e^x(x\,e^{-x}y')'$$

must be considered on $L^2(-\infty, 0; e^{-x}) \oplus L^2(0, \infty; e^{-x})$. It is limit-point at $\pm\infty$, limit-circle at $0\pm$. Hence $m = 0$, $n = 1$, $p = 1$, $q = 0$. $J_1 = (-\infty, 0)$, $K_1(0, \infty)$.

With $\lambda = 0$, we choose

$$\theta = 1, \quad \phi = \int_1^x (e^\xi/\xi)\,d\xi, \quad x > 0,$$

$$\theta = 1, \quad \phi = \int_{-1}^x (e^\xi/\xi)\,d\xi, \quad x < 0,$$

to define boundary conditions.

$$B(y) = (Q_1(y, \theta), Q_1(y, \phi), R_1(y, \theta), R_1(y, \phi))^T$$

and

$$\mathbf{J} = \begin{pmatrix} J & 0 \\ 0 & -J \end{pmatrix}.$$

Then if $(P^*, Q^*)\binom{M}{N} = \mathbf{J}$,

$$\langle L_M y, z\rangle - \langle y, L_M z\rangle = [PB(z)]^*[MB(y)] + [QB(z)]^*[NB(y)],$$

and self-adjointness occurs when $M\mathbf{J}M^* = 0$. For example,

$$M = \begin{pmatrix} 1 & -2 & 3 & 4 \\ 1 & 0 & 1 & 2 \end{pmatrix}$$

generates a mixed self-adjoint operator on $L^2(-\infty, 0; e^{-x}) \oplus L^2(0, \infty; e^{-x})$.

REFERENCES

[1] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] W. N. EVERITT AND A. ZETTL, *Sturm-Liouville differential operators in direct sum spaces*, Rocky Mountain J. Math., 16 (1986), pp. 497-516.

[3] C. T. FULTON, *Parametrization of Titchmarsh's $m(\lambda)$-functions in the limit-circle case*, Trans. Amer. Math. Soc., 229 (1977), pp. 51-63.

[4] A. M. KRALL, *Applied Analysis*, D. Reidel, Dordrecht, the Netherlands, 1986.

[5] A. M. KRALL AND A. ZETTL, *Singular self-adjoint Sturm-Liouville problems*, submitted.

[6] L. L. LITTLEJOHN AND A. M. KRALL, *Orthogonal polynomial and singular Sturm-Liouville systems*, I, Rocky Mountain J. Math., 16 (1986), pp. 435-379.

[7] ———, *Orthogonal polynomials and singular Sturm-Liouville systems*, II, submitted.

[8] M. A. NAIMARK, *Linear Differential Operators*, I *and* II, F. Ungar, New York, 1967.

[9] E. C. TITCHMARSH, *Eigenfunction Expansions*, Oxford University Press, Oxford, London, 1962.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

011

ASYMPTOTIC SOLUTIONS OF A HAMILTONIAN SYSTEM IN INTERVALS WITH SEVERAL TURNING POINTS*

HARRY GINGOLD† AND PO-FANG HSIEH‡

Abstract. The global asymptotic decomposition of a Hamiltonian system $i\varepsilon Y' = H(x)Y$, as $\varepsilon \to 0^+$, is studied. Here $H(x)$ is a Hermitian matrix analytic on $I = [a, b]$ where a could be $-\infty$ and b could be ∞. Furthermore, I may contain several turning points with various orders of this system. A complete decomposition for the asymptotic solutions is also given.

Key words. global complete asymptotic decomposition, Hamiltonian system, several turning points, asymptotic solutions

AMS(MOS) subject classifications. primary 34E20; secondary 34E15

1. Introduction. In the study of a given ordinary differential system, the process of reducing it to a simpler system by an analytic transformation is an important first step. This is also true for a system of linear equations depending on a parameter in a singular way, such as

$$\varepsilon Y' = A(x, \varepsilon)Y, \qquad ' = d/dx, \tag{1.1}$$

where ε is a small real parameter and Y and A are $n \times n$ matrices. If the eigenvalues of the leading coefficient $A(x, 0)$ coalesce on its domain of definition, the simplification process at the coalescing points, as $\varepsilon \to 0$, is usually valid only in a small neighborhood of each of such points. The points where the eigenvalues of $A(x, 0)$ coalesce are called "turning points" or "transition points" of the system (e.g., see Hsieh [7], Sibuya [14]-[16], and Wasow [19]-[21]). Thus, it is necessary to find the connection formulas if we want to study the global behavior of the system (e.g., see McHugh [9], Olver [11], [12], Sibuya [17], Turrittin [18], Wasow [20], [21], and references therein).

The purpose of this paper is to provide a *global* complete asymptotic decomposition, as $\varepsilon \to 0^+$, of the matrix Hamiltonian system

$$i\varepsilon Y' = H(x)Y, \qquad i = \sqrt{-1}, \tag{1.2}$$

where x is a real variable, ε is a parameter given in $G_{\varepsilon_0} = (0, \varepsilon_0)$, Y and $H(x)$ are $n \times n$ matrices; $H(x)$ is a Hermitian matrix analytic on $I = [a, b]$. Here a could be $-\infty$ and b could be ∞. Furthermore, the interval I may contain several (but finitely many) turning points with various (finite) orders. Such differential systems are relevant to problems in quantum mechanics. The methods presented here and in previous articles (Gingold [2] and Gingold and Hsieh [5], [6]) assisted in constructing a counterexample to the adiabatic approximation theorem in quantum mechanics (see Gingold [3]). This theorem can be found in textbooks on quantum mechanics, such as Liboff [8] and Messiah [10]. The *complete asymptotic decomposition* of solutions of (1.2) to be given here is instrumental in the proof of a modified so-called adiabatic hypothesis. Moreover, it sheds more light on the relation between transition and

* Received by the editors January 14, 1987; accepted for publication (in revised form) October 17, 1987.

† Department of Mathematics, West Virginia University, Morgantown, West Virginia 26506. The work of this author was partially supported by National Aeronautics and Space Administration research grant NAG-1-741.

‡ Department of Mathematics and Statistics, Western Michigan University, Kalamazoo, Michigan 49008-5152. The work of this author was partially supported by a Faculty Research Fellowship, Western Michigan University.

degeneracy in time-dependent self-adjoint Hamiltonian systems (see Gingold [2]–[4] and Gingold and Hsieh [6]).

2. Main theorems. In order to formulate our theorems, let $\{\lambda_1(x), \lambda_2(x), \cdots, \lambda_\sigma(x)\}$ be eigenvalues of $H(x)$ with multiplicities $n_1, n_2, \cdots, n_\sigma$ $(n_1+n_2+\cdots+n_\sigma=n)$, respectively. Moreover,

$$\lambda_j(x) \not\equiv \lambda_k(x), \quad (j \neq k) \quad \text{for } x \text{ on } I. \tag{2.1}$$

The point $x=x_0$ on I such that $\lambda_j(x_0)=\lambda_k(x_0)$ for some indices j and k $(j\neq k)$ is called a "turning point," or a "transition point," of the system (1.2). The order of zero of $\lambda_j(x)-\lambda_k(x)$ at a turning point $x=x_0$ is called the order of that turning point with respect to the pair of eigenvalues (λ_j, λ_k). It is allowed that I may contain several turning points of (1.2) with various orders.

It is noteworthy that the eigenvalues $\{\lambda_1(x), \lambda_2(x), \cdots, \lambda_\sigma(x)\}$ of $H(x)$ are real and analytic on I. By a theorem of linear algebra due to Rellich [13] there exists a unitary matrix $U(x)$ nonsingular and analytic on I such that

$$U^{-1}(x)H(x)U(x)=\lambda_1(x)I_{n_1}\oplus\lambda_2(x)I_{n_2}\oplus\cdots\oplus\lambda_\sigma(x)I_{n_\sigma}, \tag{2.2}$$

I_j: j by j identity matrix.

Let

$$Y=U(x)V; \tag{2.3}$$

then V satisfies the following equation:

$$i\varepsilon V'=\{U^{-1}(x)H(x)U(x)-i\varepsilon U^{-1}(x)U'(x)\}V. \tag{2.4}$$

Before we state our theorems, let

$$U^{-1}(x)H(x)U(x)=\Lambda_1(x)\oplus\Lambda_2(x), \tag{2.5}$$

where

$$\Lambda_1(x)=\lambda_1(x)I_{n_1}, \qquad \Lambda_2(x)=\lambda_2(x)I_{n_2}\oplus\cdots\oplus\lambda_\sigma(x)I_{n_\sigma}. \tag{2.6}$$

Also denote the partition of the second part of the coefficient of (2.4) according to that of (2.5)

$$-iU^{-1}(x)U'(x)=\begin{bmatrix} R_{11}(x) & R_{12}(x) \\ R_{21}(x) & R_{22}(x) \end{bmatrix}. \tag{2.7}$$

Namely, R_{11} is $n_1\times n_1$, R_{12} is $n_1\times(n-n_1)$, R_{21} is $(n-n_1)\times n_1$, and R_{22} is $(n-n_1)\times(n-n_1)$. Also, let

$$R(x)=\begin{bmatrix} 0 & R_{12}(x) \\ R_{21}(x) & 0 \end{bmatrix} \tag{2.8}$$

and

$$B_1(x,\varepsilon)=\Lambda_1(x)+\varepsilon R_{11}(x), \qquad B_2(x,\varepsilon)=\Lambda_2(x)+\varepsilon R_{22}(x). \tag{2.9}$$

Thus, (2.4) can be rewritten as

$$i\varepsilon V'=\{[B_1(x,\varepsilon)\oplus B_2(x,\varepsilon)]+\varepsilon R(x)\}V. \tag{2.10}$$

We shall prove the following theorem.

THEOREM 1. *There exist a positive constant* ε_1 $(0<\varepsilon_1\leqq\varepsilon_0)$ *and an* $n\times n$ *matrix* $P(x,\varepsilon)$ *analytic on* $I\times G_{\varepsilon_1}$, *such that the transformation*

$$V=W[I_n+P(x,\varepsilon)] \tag{2.11}$$

reduces (2.10) *to*

$$i\varepsilon W'=\{B_1(x,\varepsilon)\oplus B_2(x,\varepsilon)\}W, \tag{2.12}$$

where $B_j(x,\varepsilon)$ *are the matrices given by* (2.9).

Furthermore, $P(x,\varepsilon)$ *satisfies*

$$\lim P(x,\varepsilon)=0 \quad \text{as } \varepsilon \text{ tends to } 0^+. \tag{2.13}$$

We shall prove this theorem in §§ 3–5.

Applying Theorem 1 $(\sigma-1)$ times, we can get the following theorem.

THEOREM 2. *There exist a positive constant* ε_2 $(0<\varepsilon_2\leqq\varepsilon_0)$ *and an* $n\times n$ *matrix* $\hat{P}(x,\varepsilon)$ *analytic on* $I\times G_{\varepsilon_2}$, *such that the transformation*

$$V=W[I_n+\hat{P}(x,\varepsilon)] \tag{2.14}$$

reduces (2.4) *to*

$$i\varepsilon W'=\{\hat{B}_1(x,\varepsilon)\oplus\hat{B}_2(x,\varepsilon)\oplus\cdots\oplus\hat{B}_\sigma(x,\varepsilon)\}W \tag{2.15}$$

where $\hat{B}_j(x,\varepsilon)$ *are* $n_j\times n_j$ *matrices* $(j=1,2,\cdots,\sigma)$, *analytic in* $I\times G_{\varepsilon_2}$, *satisfying*

$$\hat{B}_j(x,0)=\lambda_j(x)I_{n_j} \qquad (j=1,2,\cdots,\sigma). \tag{2.16}$$

Furthermore, $\hat{P}(x,\varepsilon)$ *satisfies*

$$\lim \hat{P}(x,\varepsilon)=0 \quad \text{as } \varepsilon \text{ tends to } 0^+. \tag{2.17}$$

By the result of Theorem 2, we can obtain the following theorem.

THEOREM 3. *Let* α *be a fixed point on I. The system* (1.2) *has a fundamental matrix*

$$Y(x,\alpha,\varepsilon)=U(x)E(x,\alpha,\varepsilon)\Psi(x,\alpha)[I_n+\hat{P}(x,\varepsilon)]C \tag{2.18}$$

where $U(x)$ *is the unitary matrix given in* (2.2),

$$E(x,\alpha,\varepsilon)=E_1(x,\alpha,\varepsilon)\oplus E_2(x,\alpha,\varepsilon)\oplus\cdots\oplus E_\sigma(x,\alpha,\varepsilon) \tag{2.19}$$

with

$$E_j(x,\alpha,\varepsilon)=\exp\left\{-i\varepsilon^{-1}\int_\alpha^x\lambda_j(s)\,ds\right\}I_{n_j}, \qquad j=1,2,\cdots,\sigma, \tag{2.20}$$

$$\Psi(x,\alpha)=\Psi_1(x,\alpha)\oplus\Psi_2(x,\alpha)\oplus\cdots\oplus\Psi_\sigma(x,\alpha); \tag{2.21}$$

$\Psi_j(x,\alpha)$ *is an* $n_j\times n_j$ *unitary analytic matrix function on* $I\times G_{\varepsilon_2}$, $(j=1,2,\cdots,\sigma)$ $\hat{P}(x,\varepsilon)$ *is the matrix given in Theorem* 2, *and* C *is a suitable* $n\times n$ *constant matrix.*

This is a theorem pointed out in Gingold [3]. Its proof will be given in § 6.

Remarks. (1) Unlike other decomposition methods (e.g. Sibuya [14]–[16] and Wasow [19]) where the coefficients of the simplified equation are computed in the process, we can tell the coefficients $B_1(x,\varepsilon)$ and $B_2(x,\varepsilon)$ of the simplified equation (2.12), i.e., those given by (2.9), from the original equation (2.4). Similarly, $\hat{B}_j(x,\varepsilon)$, $(j=1,2,\cdots,\sigma)$ in (2.15) are the corresponding block-diagonal entries of (2.4). As is to be seen in § 6, $E_j(x,\alpha,\varepsilon)$ and $\Psi_j(x,\alpha)$ $(j=1,2,\cdots,\sigma)$ can be obtained directly from solving the differential equations involving the corresponding block-diagonal entries of (2.4).

(2) Unlike $U(x)$, which is obtained by finite algorithms from the given system (1.2), $\Psi_j(x, \alpha)$ are obtained as solutions of certain differential equations, or as a set of infinite series.

3. Preliminary reduction. In order to prove Theorem 1, let $Z(x, \alpha, \varepsilon)$ denote the fundamental matrix of the simplified system (2.12) satisfying

$$i\varepsilon Z' = \{B_1(x, \varepsilon) \oplus B_2(x, \varepsilon)\} Z, \qquad Z(\alpha, \alpha, \varepsilon) = I_n, \tag{3.1}$$

where α is a fixed point on I.

From (2.10), (2.11), and (2.12), we know that $P(x, \varepsilon)$ should satisfy the differential equation

$$iP' = Z^{-1} R Z (I_n + P), \tag{3.2}$$

or, equivalently, the integral equation

$$P(x, \varepsilon) = -i \int_\alpha^x Z^{-1}(s, \alpha, \varepsilon) R(s) Z(s, \alpha, \varepsilon)[I_n + P(s, \varepsilon)]\, ds. \tag{3.3}$$

Define the integral operator

$$MP = -i \int_\alpha^x Z^{-1}(s, \alpha, \varepsilon) R(s) Z(s, \alpha, \varepsilon) P(s, \varepsilon)\, ds, \tag{3.4}$$

and let

$$P_0 = MI_n. \tag{3.5}$$

Then, the integral equation (3.3) can be rewritten as

$$P = P_0 + MP. \tag{3.6}$$

In order to have a better estimate of the kernel of this integral equation, we take the second iteration of (3.6), namely,

$$P = P_0 + M^2 I_n + M^2 P, \tag{3.7}$$

where

$$\begin{aligned} M^2 P = & \int_\alpha^x Z^{-1}(s, \alpha, \varepsilon) R(s) Z(s, \alpha, \varepsilon) \\ & \cdot \left\{ \int_\alpha^s Z^{-1}(t, \alpha, \varepsilon) R(t) Z(t, \alpha, \varepsilon) P(t, \varepsilon)\, dt \right\} ds, \end{aligned} \tag{3.8}$$

or, by the change of order of integration,

$$M^2 P = \int_\alpha^x \left\{ \int_t^x Z^{-1}(s, \alpha, \varepsilon) R(s) Z(s, \alpha, \varepsilon)\, ds \right\} Z^{-1}(t, \alpha, \varepsilon) R(t) Z(t, \alpha, \varepsilon) P(t, \varepsilon)\, dt. \tag{3.9}$$

Put

$$Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}, \qquad P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}; \tag{3.10}$$

then

$$Z^{-1} R Z = \begin{bmatrix} 0 & Z_1^{-1} R_{12} Z_2 \\ Z_2^{-1} R_{21} Z_1 & 0 \end{bmatrix} \tag{3.11}$$

and

$$Z^{-1}RZP=\begin{bmatrix} Z_1^{-1}R_{12}Z_2P_{21} & Z_1^{-1}R_{12}Z_2P_{22} \\ Z_2^{-1}R_{21}Z_1P_{11} & Z_2^{-1}R_{21}Z_1P_{12}\end{bmatrix}. \tag{3.12}$$

Thus, by (3.7), (3.9), (3.11), and (3.12), we have the following decoupled equations for P_{jk}, $(j, k=1, 2)$:

$$\begin{aligned} P_{11}(x,\varepsilon)=&\int_\alpha^x\left\{\int_t^x Z_1^{-1}(s,\alpha,\varepsilon)R_{12}(s)Z_2(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z_2^{-1}(t,\alpha,\varepsilon)R_{21}(t)Z_1(t,\alpha,\varepsilon)\,dt\\ &+\int_\alpha^x\left\{\int_t^x Z_1^{-1}(s,\alpha,\varepsilon)R_{12}(s)Z_2(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z_2^{-1}(t,\alpha,\varepsilon)R_{21}(t)Z_1(t,\alpha,\varepsilon)P_{11}(t,\varepsilon)\,dt, \end{aligned} \tag{3.13}$$

$$\begin{aligned} P_{12}(x,\varepsilon)=&\int_\alpha^x Z_1^{-1}(s,\alpha,\varepsilon)R_{12}(s)Z_2(s,\alpha,\varepsilon)\,ds\\ &+\int_\alpha^x\left\{\int_t^x Z_1^{-1}(s,\alpha,\varepsilon)R_{12}(s)Z_2(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z_2^{-1}(t,\alpha,\varepsilon)R_{21}(t)Z_1(t,\alpha,\varepsilon)P_{12}(t,\varepsilon)\,dt, \end{aligned} \tag{3.14}$$

$$\begin{aligned} P_{21}(x,\varepsilon)=&\int_\alpha^x Z_2^{-1}(s,\alpha,\varepsilon)R_{21}(s)Z_1(s,\alpha,\varepsilon)\,ds\\ &+\int_\alpha^x\left\{\int_t^x Z_2^{-1}(s,\alpha,\varepsilon)R_{21}(s)Z_1(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z_1^{-1}(t,\alpha,\varepsilon)R_{12}(t)Z_2(t,\alpha,\varepsilon)P_{21}(t,\varepsilon)\,dt, \end{aligned} \tag{3.15}$$

$$\begin{aligned} P_{22}(x,\varepsilon)=&\int_\alpha^s\left\{\int_t^x Z_2^{-1}(s,\alpha,\varepsilon)R_{21}(s)Z_1(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z_1^{-1}(t,\alpha,\varepsilon)R_{12}(t)Z_2(t,\alpha,\varepsilon)\,dt\\ &+\int_\alpha^x\left\{\int_t^x Z_2^{-1}(s,\alpha,\varepsilon)R_{21}(s)Z_1(s,\alpha,\varepsilon)\,ds\right\}\\ &\cdot Z_1^{-1}(t,\alpha,\varepsilon)R_{12}(t)Z_2(t,\alpha,\varepsilon)P_{22}(t,\varepsilon)\,dt. \end{aligned} \tag{3.16}$$

We shall prove that there exist the solutions P_{jk} of (3.13)-(3.16), respectively, satisfying the required properties described in Theorem 1. It is noteworthy that these decoupled equations (3.13)-(3.16) are analogous to the scalar equations obtained for a two-dimensional system used in Gingold [2].

4. A fundamental lemma. In order to prove the existence of the solution P_{jk} of (3.13)-(3.16) we must have certain estimates of the magnitude of the kernels in their respective integrals. For this purpose, we must invoke a lemma of ours proved recently in [6].

LEMMA 1. *Let there be given an integral expression*

$$J(a,b)=\int_a^b r(s,\varepsilon)\exp\left\{i\varepsilon^{-1}\int_\alpha^s p(\eta)\,d\eta\right\}ds, \qquad a\leqq\alpha\leqq b, \tag{4.1}$$

where $r(x, \varepsilon)$ is in the class $C^1(I \times \bar{G}_{\varepsilon_0})$ and $p(x)$ is real analytic on $I \times \bar{G}_{\varepsilon_0}$. Furthermore, assume the following:

(i) *$p(x)$ may vanish at some points of I, but it is not identically zero in I.*

(ii) *$r(x, \varepsilon)$ is uniformly bounded on $I \times \bar{G}_{\varepsilon_0}$.*

(iii) *Both $r(x, \varepsilon)$ and $r'(x, \varepsilon)$ are absolutely integrable over I for ε in G_{ε_0}.*

Then, there exist positive constants $\hat{K}$, γ, and $\hat{\varepsilon}$ $(0 < \gamma < 1,\ 0 < \hat{\varepsilon} \leqq \varepsilon_0)$ such that

$$|J(a, b)| \leqq \hat{K}\varepsilon^{\gamma} \quad \text{for } \varepsilon \text{ in } G_{\hat{\varepsilon}}. \tag{4.2}$$

Remark. As defined in [6], the zero points of $p(x)$ on I are called the turning points of the integral expression $J(a, b)$.

To simplify the notation, let

$$\hat{n}_1 = n_1, \qquad \hat{n}_2 = n - n_1. \tag{4.3}$$

In the notation of (3.10) and (3.9), $Z_1(x, \alpha, \varepsilon)$ and $Z_2(x, \alpha, \varepsilon)$ satisfy the following initial value problem:

$$i\varepsilon Z_j' = [\Lambda_j(x) + \varepsilon R_{jj}(x)]Z_j, \quad Z_j(\alpha, \alpha, \varepsilon) = I_{\hat{n}_j}, \quad j = 1, 2. \tag{4.4}$$

Put

$$Z_j(x, \alpha, \varepsilon) = \Phi_{1j}(x, \alpha, \varepsilon)\Phi_{2j}(x, \alpha, \varepsilon), \qquad j = 1, 2, \tag{4.5}$$

where $\Phi_{1j}(x, \alpha, \varepsilon)$ satisfies the equation

$$i\varepsilon \Phi_{1j}' = \Lambda_j(x)\Phi_{1j}, \quad \Phi_{1j}(\alpha, \alpha, \varepsilon) = I_{\hat{n}_j}, \quad j = 1, 2. \tag{4.6}$$

Since $\Lambda_j(x)$, $(j = 1, 2)$ is diagonal in the form of (2.6), we have

$$\Phi_{1j}(x, \alpha, \varepsilon) = \exp\left\{-i\varepsilon^{-1}\int_{\alpha}^{x} \Lambda_j(s)\, ds\right\}, \qquad j = 1, 2. \tag{4.7}$$

Furthermore, $\Phi_{2j}(x, \alpha, \varepsilon)$ satisfies the conditions

$$i\Phi_{2j}' = [\Phi_{1j}^{-1}(x, \alpha, \varepsilon)R_{jj}(x)\Phi_{1j}(x, \alpha, \varepsilon)]\Phi_{2j}, \qquad \Phi_{2j}(\alpha, \alpha, \varepsilon) = I_{\hat{n}_j}. \tag{4.8}$$

By the assumptions of $\Lambda_j(x)$ and $R_{jj}(x)$, we know that $\Phi_{2j}(x, \alpha, \varepsilon)$ are in the class of C^1 on $I \times I \times G_{\varepsilon_0}$ (e.g., see Coddington and Levinson [1]). Thus, the entries of $\Phi_{2j}'(x, \alpha, \varepsilon)$ are also absolutely integrable $(j = 1, 2)$.

Now, by (4.5), let

$$\begin{aligned}(r_{(jk)\mu\nu}) &:= Z_j^{-1}R_{jk}Z_k = \Phi_{2j}^{-1}\Phi_{1j}^{-1}R_{jk}\Phi_{1k}\Phi_{2k}, \\ &\mu = 1, 2, \cdots, \hat{n}_j, \quad \nu = 1, 2, \cdots, \hat{n}_k \quad (j = 1, 2).\end{aligned} \tag{4.9}$$

Since Φ_{1j} are diagonal and are in the form of (4.7), we know that $r_{(jk)\mu\nu}$ are finite sums of the terms of the form

$$r_{(jk)\mu\nu} = \sum_{\rho=1}^{\hat{n}_j \times \hat{n}_k} \tilde{r}_{(jk)\mu\nu\rho}(x, \alpha, \varepsilon) \exp\left\{i\varepsilon^{-1}\int_{\alpha}^{x} p_{(jk)\mu\nu\rho}(s)\, ds\right\}, \tag{4.10}$$

where each of $p_{(jk)\mu\nu\rho}(x)$ is one of the difference of the diagonal elements of $\Lambda_1(x)$ and those of $\Lambda_2(x)$, or vice versa. $\tilde{r}_{(jk)\mu\nu\rho}(x, \alpha, \varepsilon)$ are products of the elements of Φ_{2j}^{-1}, R_{jk}, and Φ_{2k}. Thus $\tilde{r}_{(jk)\mu\nu\rho}(x, \alpha, \varepsilon)$ are analytic and bounded for $I \times G_{\varepsilon_0}$. Furthermore, $\tilde{r}_{(jk)\mu\nu\rho}(x, \alpha, \varepsilon)$ and their derivatives, with respect to x, are absolutely integrable on I. Therefore, every element of the matrix $Z_j^{-1}R_{jk}Z_k$ satisfies the assumptions of Lemma 1.

Now, let $\|\cdot\|$ denote a suitable norm of a matrix. By the use of Lemma 1 and the discussion above, we have the following lemma.

LEMMA 2. *For each pair* (j, k) $(j, k = 1, 2)$ *there exist positive constants* K_{jk}, d_{jk}, *and* ε_{jk} $(0 < d_{jk} < 1,\ 0 < \varepsilon_{jk} \leqq \varepsilon_0)$ *such that*

$$\left\| \int_\alpha^x Z_j^{-1}(s, \alpha, \varepsilon) R_{jk}(s) Z_k(s, \alpha, \varepsilon)\, ds \right\| < K_{jk} \varepsilon^{d_{jk}} \tag{4.11}$$

for all α *and* x *on* I, *and* $0 < \varepsilon < \varepsilon_{jk}$.

5. Completion of the proof of Theorem 1. By applying Lemma 2 to (3.13)-(3.16), we know that for each pair (j, k) $(j, k = 1, 2)$ there exist a positive constant $\tilde{K}_{jk}$ and a positive function $\varphi_{jk}(\varepsilon)$ such that

$$\lim \varphi_{jk}(\varepsilon) = 0 \quad \text{as } \varepsilon \text{ tends to } 0^+, \tag{5.1}$$

and

$$\|P_{jk}\| \leqq \varphi_{jk}(\varepsilon) + \tilde{K}_{jk} \varepsilon^{d_{jk}} \|P_{jk}\| \quad \text{for } x \text{ on } I \text{ and } 0 < \varepsilon < \varepsilon_{jk}. \tag{5.2}$$

Now, choose ε_1 sufficiently small such that

$$\varepsilon_1 \leqq \min \{\varepsilon_{jk} \,|\, j, k = 1, 2\} \tag{5.3}$$

and

$$\tilde{K}_{jk} \varepsilon_1^{d_{jk}} < 1 \qquad \text{for } j, k = 1, 2. \tag{5.4}$$

Thus, each of the equations (3.13)-(3.16) defines a contraction mapping for x on I and ε in G_{ε_1}. Therefore their solutions P_{jk} exist and, furthermore,

$$\|P_{jk}\| \leqq \varphi_{jk}(\varepsilon)/(1 - \tilde{K}_{jk} \varepsilon^{d_{jk}}), \qquad j, k = 1, 2. \tag{5.5}$$

Hence, each of P_{jk} satisfies the relation (2.13), and Theorem 1 is proved.

6. Proof of Theorem 3. In order to prove Theorem 3, let

$$-iU^{-1}(x) U'(x) = \begin{bmatrix} \hat{R}_{11}(x) & \hat{R}_{12}(x) & \cdots & \hat{R}_{1\sigma}(x) \\ \hat{R}_{21}(x) & \hat{R}_{22}(x) & \cdots & \hat{R}_{2\sigma}(x) \\ \vdots & \vdots & & \vdots \\ \hat{R}_{\sigma 1}(x) & \hat{R}_{\sigma 2}(x) & \cdots & \hat{R}_{\sigma\sigma}(x) \end{bmatrix} \tag{6.1}$$

where $\hat{R}_{jk}(x)$ are $n_j \times n_k$ matrices $(j, k = 1, 2, \cdots, \sigma)$. Then,

$$\hat{B}_j(x, \varepsilon) = \lambda_j(x) I_{n_j} + \varepsilon \hat{R}_{jj}(x). \tag{6.2}$$

Also, let $W_j(x, \alpha, \varepsilon)$ be the solution of

$$i\varepsilon W_j' = \hat{B}_j(x, \varepsilon) W_j, \qquad W_j(\alpha, \alpha, \varepsilon) = I_{n_j}. \tag{6.3}$$

Then the fundamental matrix (2.15) is given by

$$W(x, \alpha, \varepsilon) = \{W_1(x, \alpha, \varepsilon) \oplus W_2(x, \alpha, \varepsilon) \oplus \cdots \oplus W_\sigma(x, \alpha, \varepsilon)\} G, \tag{6.4}$$

where G is an $n \times n$ constant matrix. Similar to (4.5), let

$$W_j(x, \alpha, \varepsilon) = E_j(x, \alpha, \varepsilon) \Psi_j(x, \alpha), \tag{6.5}$$

where $E_j(x, \alpha, \varepsilon)$ is given by (2.20). Namely, it satisfies

$$i\varepsilon E_j' = \lambda_j(x) E_j, \qquad E_j(\alpha, \alpha, \varepsilon) = I_{n_j}. \tag{6.6}$$

Since $E_j(x, \alpha, \varepsilon)$ is in the form of (2.20), $\Psi_j(x, \alpha)$ satisfies the equation

$$i\Psi_j' = \hat{R}_{jj}(x) \Psi_j, \qquad \Psi_j(\alpha, \alpha) = I_{n_j}. \tag{6.7}$$

Observe first that, for an analytic unitary matrix $U(x)$, $U^{-1}(x)U'(x)$ is anti-Hermitian. By the facts that the independent variable x is real and the operations of differentiation and conjugate transposition are commutative, we have

$$\begin{aligned}[U^{-1}U']^* &= [U^{-1}\{-U(U^{-1})'U\}]^* = -[(U^{-1})'U]^* \\ &= -U^*[(U^{-1})']^* = -U^{-1}U'.\end{aligned} \tag{6.8}$$

Thus, by (6.1), $\hat{R}_{jj}(x)$ $(j=1,2,\cdots,\sigma)$ are Hermitian. Therefore, by (6.7), Ψ_j $(j=1,2,\cdots,\sigma)$ are unitary. Hence, Theorem 3 is proved.

7. Concluding remark. By repeating the iterations of (3.6), we have the infinite series expansion for the matrix

$$P=\sum_{\nu=0}^{\infty} M^{\nu}I_n. \tag{7.1}$$

As pointed out in [2], by a similar estimate as that given in Lemma 2, there exist two constants K and d $(0<d<1)$, such that

$$\|M^{\nu}I_n\| \leqq K^{\nu}\varepsilon^{\nu d}, \qquad \nu=1,2,3,\cdots \tag{7.2}$$

for all x on I. Thus the series (7.1) represents a generalized asymptotic expansion. Moreover, it is possible to extract from the series of (7.1), an asymptotic series in the sense of Poincaré wherever it exists (compare with [2]).

Note added in proof. By (2.4), (2.14), (2.15), (2.16), and (6.2), the matrix $I_n+\hat{P}(x,\varepsilon)$ satisfies a differential equation similar to (6.7). Thus, $I_n+\hat{P}(x,\varepsilon)$ is unitary.

REFERENCES

[1] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[2] H. GINGOLD, *An asymptotic decomposition method applied to multiturning point problems*, SIAM J. Math. Anal., 16 (1985), pp. 7-27.

[3] ———, *A counterexample to the adiabatic approximation theorem in quantum mechanics*, West Virginia University, preprint.

[4] ———, *In general, the less degeneracy the less transition. A principle for time dependent Hamiltonian systems in quantum mechanics*, J. Math. Phys., 28 (1987), pp. 2400-2406.

[5] H. GINGOLD AND P. F. HSIEH, *Global simplification of a singularly perturbed almost diagonal system*, SIAM J. Math. Anal., 17 (1986), pp. 7-18.

[6] ———, *Global approximation of perturbed Hamiltonian differential equations with several turning points*, SIAM J. Math. Anal., 18 (1987), pp. 1275-1293.

[7] P. F. HSIEH, *A turning point problem for a system of linear ordinary differential equations of the third order*, Arch. Rational Mech. Anal., 19 (1965), pp. 117-148.

[8] R. L. LIBOFF, *Introductory Quantum Mechanics*, Holden-Day, San Francisco, 1980.

[9] J. A. M. MCHUGH, *An historical survey of ordinary linear differential equations with a large parameter and turning points*, Arch. Hist. Exact Sci., 7 (1971), pp. 277-324.

[10] A. MESSIAH, *Quantum Mechanics, Vol.* 2, Interscience, New York, 1961.

[11] F. W. J. OLVER, *Connection formulas for second order differential equations with multiple turning points*, SIAM J. Math. Anal., 8 (1977), pp. 127-154.

[12] ———, *Connection formulas for second order differential equations having an arbitrary number of turning points of arbitrary multiplicities*, SIAM J. Math. Anal., 8 (1977), pp. 673-700.

[13] F. RELLICH, *Störungstheorie der Spektralzerlegung, I Mitteilung*, Math. Ann., 113 (1936), pp. 600-619.

[14] Y. SIBUYA, *Simplification of a system of linear ordinary differential equations about a singular point*, Funkcial Ekvac., 4 (1962), pp. 29-56.

[15] ———, *Asymptotic solutions of a system of linear ordinary differential equations containing a parameter*, Funkcial Ekvac., 4 (1962), pp. 115-139.

[16] Y. SIBUYA, *Simplification of a linear ordinary differential equation of the nth order at a turning point*, Arch. Rational Mech. Anal., 13 (1963), pp. 206–221.

[17] ———, *Global Theory of a Second Order Linear Ordinary Differential Equation with a Polynomial Coefficient*, North-Holland–American Elsevier, Amsterdam, New York, 1975.

[18] H. L. TURRITTIN, *Solvable related equations pertaining to turning point problems*, in Asymptotic Solutions of Differential Equations and Their Applications, C. H. Wilcox, ed., John Wiley, New York, 1964, pp. 27–52.

[19] W. WASOW, *Asymptotic Expansions for Ordinary Differential Equations*, John Wiley, New York, 1965.

[20] ———, *The central connection problem at turning points of linear differential equations*, Comm. Math. Helvetici, 46 (1971), pp. 65–86.

[21] ———, *Linear Turning Point Theory*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

012

EXTREMAL PROBLEMS FOR EIGENVALUE FUNCTIONALS II*

DAVID C. BARNES†

Abstract. This paper is actually the fourth in a series of works [*SIAM J. Math. Anal.*, 16(1985), pp. 341-357, 1284-1294; 18(1987), pp. 933-940] whose overall purpose is to develop some variational and approximation theory for eigenvalue functionals. These functionals are defined by the eigenvalues of differential operators, of the general form $\mathcal{L}(\cdot) = \lambda M(\cdot)$. We assume the operators $\mathcal{L}(\cdot)$ and $M(\cdot)$ have coefficients that depend on some real valued function (or functions) $\rho(x)$ taken from a class C. When subject to appropriate boundary conditions, the eigenvalues λ become real valued functionals of ρ. The functionals are defined on the class C, and are denoted by $\lambda_n(\rho)$. In this work we show a new way to apply the classical theory of variational calculus to the problem of maximizing or minimizing a function of two (or more) such eigenvalues $\Phi(\lambda_n(\rho), \mu_m(\rho))$. We then give an extensive application of the example $\Phi(u, v) = \min\{u, v\}$ to a three-dimensional version of the strongest column problem.

Key words. strongest column, eigenvalue, extremal problem

AMS(MOS) subject classifications. 35P15, 49A99

1. The general theory of extremals for functions of eigenvalues. Consider a vibrating string having density function $\rho(x)$ with fixed end points. Its characteristic frequencies of vibration are determined by the eigenvalues $\lambda = \lambda(\rho)$ of the system

$$y'' + \lambda\rho(x)y = 0, \qquad y(0) = y(l) = 0, \qquad 0 \le x \le l. \tag{1}$$

There will be an infinite sequence of eigenvalues $\lambda_1(\rho) \le \lambda_2(\rho) \le \lambda_3(\rho) \le \cdots$. Several works [21],[13],[9] have considered the problem of maximizing and minimizing the ratio $\lambda_1(\rho)/\lambda_2(\rho)$. Theorem 1 below, with $\Phi(u, v) = u/v$, can be used to easily reproduce the extremal conditions used in those works.

Suppose a number of such strings (and perhaps some rods as well) are all vibrating together and that they each have density function $\rho_i(x)$, length l_i and eigenvalues $\lambda_n(\rho_i)$. The fundamental frequencies of vibration Λ and the total mass M of the system are now determined by

$$\Lambda = \min_i\{\lambda_1(\rho_i)\}, \qquad M = \sum_i \int_0^{l_i} \rho_i(x)\,dx.$$

One might want to minimize Λ subject to a given mass constraint. In a similar way, consider a single large load, being supported by several columns of length l_i and total volume V. The critical buckling load of each column is determined by an eigenvalue problem similar to (1). To maximize the load-carrying capacity of the structure, one would want to maximize the sum of all of the first eigenvalues for each of the systems.

We will now give a general theory of the extremals for functions of eigenvalues $\Phi(\lambda_n, \mu_m)$. We will assume, whenever necessary, that the problem under study does actually have an extremal function, denoted by $\rho^*(x)$. Methods used by Barnes [6] or Holmåker [11] could, perhaps, be used to give proofs of that assumption.

* Received by the editors March 2, 1987; accepted for publication October 10, 1987. This research was supported by the Northwest College and University Association for Science (University of Washington) under contract DE-AM06-76-RL02225 with the U.S. Department of Energy.

† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164-2930.

1.1. The necessary conditions. Consider differential operators $\mathcal{L}(\cdot)$, $M(\cdot)$ of the general form

$$\mathcal{L}(\rho, y) = \sum_{i=0}^{m} (-1)^i \left(f_i(x,\rho) y^{(i)}\right)^{(i)}, \qquad M(\rho, y) = \sum_{j=0}^{m'} (-1)^j \left(g_j(\rho, x) y^{(j)}\right)^{(j)}.$$

The operators will be linear in y but may be nonlinear in ρ. Here, the function ρ ranges through some class C, and the coefficient functions f_i and g_j depend on x as well as the function $\rho(x)$. We shall use the abbreviated notation $f_i[\rho] = f_i(x, \rho(x))$ for the various coefficients that are assumed to satisfy appropriate smoothness conditions and so on. For more details, see [20],[4].

We will be concerned with generalized eigenvalue problems of the form

$$\mathcal{L}(\rho, y) = \lambda M(\rho, y), \qquad U_p(y, \rho, \lambda) = 0, \qquad p = 1, 2, \cdots, 2m. \tag{2}$$

We will assume that $m > m' \geq 0$. The boundary conditions U_p may also depend on the values of ρ at $x = 0$ or l, as well as the eigenvalue parameter λ.

Now, suppose that two distinct eigenvalue problems of the form (2) are given, so we have four operators, $\mathcal{L}_i(\cdot)$ and $M_i(\cdot)$, each with coefficients depending on $\rho \in C$. The two eigenvalue equations are

$$\mathcal{L}_1(\rho, y) = \lambda M_1(\rho, y), \qquad \mathcal{L}_2(\rho, z) = \mu M_2(\rho, z). \tag{3}$$

Suppose that $\Phi(u, v)$ is a given function of u, v. The major problem considered in this work is the following.

Problem 1. Let C be a given class of functions $\rho(x)$ and let $\lambda_n(\rho), \mu_m(\rho)$ be eigenvalues of (3). Find that function $\rho^*(x) \in C$ that maximizes (minimizes) the functional $\Phi(\lambda_n(\rho), \mu_m(\rho))$ for all functions $\rho \in C$.

The eigenvalues for these systems are functionals defined on the class C of functions $\rho(x)$. Let ρ^* be a fixed function in C and suppose that it is smoothly imbedded into a family of functions $\rho_\epsilon(x) \in C$ so that $\rho_0(x) = \rho^*(x)$. Using methods given by Barnes [5], we obtain the following theorem.

THEOREM 1. *Given functions ρ^*, $\rho_\epsilon \in C$, let y^* and y be eigenfunctions for* (2), *corresponding to eigenvalues $\lambda(\rho^*)$ and $\lambda(\rho_\epsilon)$. Define a functional $J(\rho)$ on the class C by*

$$J(\rho) = \int_0^l \left\{ \sum_{j=0}^{m'} g_j[\rho] \left(y^{*(j)}\right)^2 - \sum_{i=0}^{m} f_i[\rho] \left(y^{*(i)}\right)^2 \right\} dx.$$

Define boundary terms $BT1$, $BT2$, and $BT3$ by the following equations:

$$\begin{gathered} (\mathcal{L}(\rho^*, y), y^*) = BT1 + (y, \mathcal{L}(\rho^*, y^*)), \\ (M(\rho^*, y), y^*) = BT2 + (y, M(\rho^*, y^*)), \\ (\mathcal{L}(\rho, y^*), y^*) - \lambda^*(M(\rho, y^*), y^*) = BT3 - J(\rho). \end{gathered}$$

Suppose that y^* is normalized so that

$$(M(\rho^*, y^*), y^*) = 1, \qquad (\mathcal{L}(\rho^*, y^*), y^*) = \lambda^*.$$

Finally, define a functional $K(\rho)$ on C by

$$K(\rho) = \lambda^* + BT1 - \lambda^* BT2 + BT3 - J(\rho).$$

Then

$$\begin{aligned}\lambda(\rho_\epsilon) &= K(\rho_\epsilon) + O(\epsilon^2)\\ &= \lambda^* + BT1 - \lambda^* BT2 + BT3 - J(\rho_\epsilon) + O(\epsilon^2).\end{aligned}$$

Thus, the two functionals $K(\rho)$ *and* $\lambda(\rho)$ *are tangent to each other at* $\rho = \rho^*$. *That is,*

$$K(\rho^*) = \lambda(\rho^*), \quad \textit{and, at } \rho = \rho^*, \quad \delta K = \delta\lambda.$$

The term $O(\epsilon^2)$ *is given by*

$$O(\epsilon^2) = (\Delta\mathcal{L}(\Delta y) - \Delta\lambda M^*(\Delta y) - \Delta\lambda\Delta M(y^*) - \lambda\Delta M(\Delta y)\,,\, y^*)$$

where $\Delta(\cdot) = (\cdot) - (\cdot)^*$.

It frequently happens that the boundary terms can be calculated in a simple form that is independent of the eigenfunction y. If that is so, then Theorem 1 shows that ρ^* is an extremal of $\lambda(\rho)$ if and only if it is also an extremal of the functional $J(\rho)$. Now $J(\rho)$ is a classical functional defined on $\mathcal{C}$, and classical methods can be used to find its extremals. This was the essential idea used in the works [3],[4]. To apply these methods to Problem 1, we first use Theorem 1 to construct the two functionals $K_i(\rho)$ associated with the eigenvalue problems (3). The solution of Problem 1 can then be found using the following theorem.

THEOREM 2. *Suppose that* $\Phi(u, v)$ *satisfies the Lipschitz condition*

$$|\Phi(u_1, v_1) - \Phi(u_2, v_2)| \le L\{\,|u_1 - u_2| + |v_1 - v_2|\,\}$$

and that y^*, z^* *are eigenfunctions of* (3) *corresponding to* $\lambda(\rho^*)$ *and* $\mu(\rho^*)$. *Let* $K_1(\rho)$ *and* $K_2(\rho)$ *be the functionals corresponding to* (3) *defined by Theorem* 1. *Then the two functionals* $\Phi(\lambda(\rho), \mu(\rho))$ *and* $\Phi(K_1(\rho), K_2(\rho))$ *are tangent to each other at* $\rho = \rho^*$. *That is,*

$$\begin{gathered}\Phi(\lambda(\rho_\epsilon), \mu(\rho_\epsilon)) = \Phi(K_1(\rho_\epsilon), K_2(\rho_\epsilon)) + O(\epsilon^2),\\ \Phi(\lambda(\rho^*), \mu(\rho^*)) = \Phi(K_1(\rho^*), K_2(\rho^*)),\\ \delta\Phi(\lambda(\rho), \mu(\rho)) = \delta\Phi(K_1(\rho), K_2(\rho)) \textit{ at } \rho = \rho^*.\end{gathered}$$

The proof of this theorem follows immediately from Theorem 1 and the inequality

$$|\Phi(\lambda(\rho_\epsilon), \mu(\rho_\epsilon)) - \Phi(K_1(\rho_\epsilon), K_2(\rho_\epsilon))| \le L\{\,|\lambda(\rho_\epsilon) - K_1(\rho_\epsilon)| + |\mu(\rho_\epsilon) - K_2(\rho_\epsilon)|\,\} = O(\epsilon^2).$$

It follows that ρ^* is an extremal of $\Phi(\lambda(\rho), \mu(\rho))$ if and only if it is also an extremal of the classical functional $\Phi(K_1(\rho), K_2(\rho))$. We may now proceed to find ρ^* using well-known methods. Incidentally, this theory requires only that the function Φ satisfy a Lipschitz condition, not that it be differentiable as other treatments of the problem have assumed. This is especially critical for the example $\Phi(u, v) = \min\{u, v\}$ considered below.

2. The strongest column in three dimensions. An interesting problem, with a long history (see, for example, [16],[8],[12],[18],[14],[2]) is to design a column having a given length and volume, so that, when subject to an axial compressive load, the critical buckling load is as large as possible. We will now investigate this problem allowing certain kinds of three-dimensional variation in its cross-sectional shape, and allowing for three-dimensional supports of the end points. Suppose the column lies on the x axis, has length l, is perpendicular to the y-z plane and, at $x = 0$, is attached to

a crankshaft using a journal bearing. The crankshaft is the y axis. Thus, the column is pinned in the x-z plane but it is clamped in the x-y plane. Suppose that it is pinned in both planes at $x = l$.

Let $\mathcal{D}_x$ be a cross section of the column, taken by a plane parallel to the y-z plane at a distance x from it. We will need the two moments of inertia $I_y(\mathcal{D}_x)$ and $I_z(\mathcal{D}_x)$ of the cross section defined by

$$I_y(\mathcal{D}_x) = \iint_{\mathcal{D}_x} z^2\,dy\,dz, \qquad I_z(\mathcal{D}_x) = \iint_{\mathcal{D}_x} y^2\,dy\,dz.$$

The critical buckling load in the x-y plane is determined by the second eigenvalue λ_2, of the system [19],[2]

$$y'' + \frac{\lambda}{I_z(\mathcal{D}_x)}y = 0, \qquad y(0) + y'(0) = 0, \quad y(l) = 0,$$

while the critical buckling load in the x-z plane is determined by the first eigenvalue μ_1, of the system

$$z'' + \frac{\mu}{I_y(\mathcal{D}_x)}z = 0, \qquad z(0) = 0, \quad z(l) = 0.$$

For now, we will only consider buckling in these two planes. In §2.4 we will consider diagonal planes.

We will suppose that, at any given x, the shape of the cross section $\mathcal{D}_x$ is determined by using some given geometrical figure, call it $\widehat{\mathcal{D}}$, such as a circle or square, and applying to $\widehat{\mathcal{D}}$ a stretching factor, $a(x)$ in the y direction, and $b(x)$ in the z direction, to obtain the cross section $\mathcal{D}_x$. So, for example, if $\widehat{\mathcal{D}}$ is a circle, then $\mathcal{D}_x$ will be an ellipse; if $\widehat{\mathcal{D}}$ is a square, then $\mathcal{D}_x$ will be a rectangle. The quantities $a(x)$, $b(x)$ are nondimensional variables.

We will assume, for convenience and without loss of generality, that the two moments of inertia of $\widehat{\mathcal{D}}$ satisfy $I_y(\widehat{\mathcal{D}}) = I_z(\widehat{\mathcal{D}}) = 1$. Indeed, if this were not true, then we could first stretch $\widehat{\mathcal{D}}$ one way or another to make it true, and then simply multiply $a(x)$ and $b(x)$ by appropriate constants and still maintain the same size and shape for the column.

We will now map $\widehat{\mathcal{D}}$ onto $\mathcal{D}_x$ using the stretching transformation $y = a(x)\widehat{y}$, $z = b(x)\widehat{z}$. We see that

$$I_y(\mathcal{D}_x) = \iint_{\mathcal{D}_x} y^2\,dy\,dz = a^3(x)b(x)\iint_{\widehat{\mathcal{D}}} \widehat{y}^2\,d\widehat{y}\,d\widehat{z} = a^3(x)b(x).$$

Similarly, $I_z(\mathcal{D}_x) = a(x)b^3(x)$. Thus, we obtain the equations

$$y'' + \frac{\lambda}{a^3(x)b(x)}y = 0, \qquad y(0) + y'(0) = y(l) = 0, \tag{4}$$

$$z'' + \frac{\mu}{a(x)b^3(x)}z = 0, \qquad z(0) = z(l) = 0. \tag{5}$$

We use the following notation: $A(x)$ is the area of $\mathcal{D}_x$, $\widehat{A}$ is the area of $\widehat{\mathcal{D}}$, V is the volume of the column, and $\widehat{V} = V/\widehat{A}$. It follows that

$$A(x) = \iint_{\mathcal{D}_x} dy\,dz = a(x)b(x)\widehat{A}, \qquad \int_0^l a(x)b(x)\,dx = \widehat{V}. \tag{6}$$

In order to avoid certain difficulties with singular points [6], we will also assume that constants H_1, H_2, H_3, H_4 are given and that the functions $a(x)$, $b(x)$ are constrained by the conditions

$$(7)\quad 0 < H_1 \le a(x) \le H_2, \qquad 0 < H_3 \le b(x) \le H_4, \qquad lH_1H_3 < \widehat{V} < lH_2H_4.$$

We let $\mathcal{C}$ denote the class of all piecewise continuous function pairs $(a(x), b(x))$ satisfying (7). The eigenvalues of (4), (5) are real valued functionals of the pair $(a(x), b(x))$, and we denote them by $\lambda_n(a,b)$, $\mu_m(a,b)$. The critical buckling load in the x-y plane is determined by $\lambda_2(a,b)$ and in the x-z plane by $\mu_1(a,b)$. If we want the strongest column, then we need to maximize the minimum of the two.

Problem 2. Let $\lambda_2(a,b)$ be the second eigenvalue of (5), and let $\mu_1(a,b)$ be the first eigenvalue of (4). Find $(a^*(x), b^*(x)) \in \mathcal{C}$, that maximizes the functional

$$\min\{\lambda_2(a,b),\ \mu_1(a,b)\}$$

over all $(a,b) \in \mathcal{C}$.

We will assume that a solution $(a^*(x), b^*(x)) \in \mathcal{C}$ of Problem 2 exists.

2.1. The necessary conditions for the strongest column. Suppose that we are given extremals $(a^*(x), b^*(x))$ for Problem 2 and let y^*, z^* be the corresponding eigenfunctions of (4), (5). To obtain the necessary conditions, we first use Theorem 1 to calculate the functionals $K_i(a,b)$ finding,

$$(8)\quad \begin{aligned} K_1(a,b) &= 2\lambda_2^* - y^{*2}(0) - \int_0^l \frac{\lambda_2^*}{a^3(x)b(x)} y^{*2}\,dx, \\ K_2(a,b) &= 2\mu_1^* - \int_0^l \frac{\mu_1^*}{a(x)b^3(x)} z^{*2}\,dx. \end{aligned}$$

We now use Theorem 2, with $\Phi(u,v) = \min\{u,v\}$ to transform Problem 2 into the following more easily understood problem.

Problem 3. With K_1, K_2 defined by (8) find $(a^*(x), b^*(x)) \in \mathcal{C}$, that maximizes the functional

$$\min\{K_1(a,b),\ K_2(a,b)\}$$

over all $(a,b) \in \mathcal{C}$.

In most cases, the extremal (a^*, b^*) will be such that $\lambda_2^* = \mu_1^*$. If not, we could shave a little mass from one side of the column and paste it onto the other, thereby increasing its overall strength. The only time this is not possible is when the constraints H_i interfere with this possibility (if, for example, $a^*(x) = H_1$ for all x). Although it could be easily done, we will not give the details of the analysis in such degenerate cases since, with any reasonable choice of H_i, a little working space would be left over. It would require consideration of various cases, in which we take either $a(x) \equiv H_i$ or $b(x) \equiv H_i$ and, one by one, optimize over the other function. Such problems have been dealt with extensively before. Thus we will assume that $K_1(a^*, b^*) = K_2(a^*, b^*)$. This relationship can now be used to transform Problem 3 into the following form.

Problem 4. Find $(a^*(x), b^*(x)) \in \mathcal{C}$, that maximizes the functional $K_1(a,b)$ over all $(a(x), b(x)) \in \mathcal{C}$ satisfying the additional constraint $K_1(a,b) - K_2(a,b) = 0$. That is,

$$2\lambda_2^* - y^{*2}(0) - 2\mu_1^* + \int_0^l \left\{ \frac{\mu_1^*}{a(x)b^3(x)} z^{*2} - \frac{\lambda_2^*}{a^3(x)b(x)} y^{*2} \right\} dx = 0.$$

This is one example of a class of rather ordinary problems in calculus of variations. Generally, they are of the following form.

Problem 5. Find $\rho^*(x)$ so that $J(\rho) = \int_0^l F_0(x, \rho(x))\,dx$ is a maximum subject to the constraints,

$$\int_0^l F_i(x, \rho(x))\,dx = V_i \quad \text{for } i = 1, 2, 3, \cdots, N \text{ and } \quad h \le \rho(x) \le H.$$

The solution of such problems can be obtained using the method of Lagrange Multipliers. See, for example, the books by Hestenes [10] or Troutman [20] for the following theorem.

THEOREM 3. *Let $F_i(x, \rho)$ be continuous and let $\rho^*(x)$ be a solution of Problem* 5. *Then there exist constants $\gamma_0 \ge 0$ and $\gamma_1, \gamma_2, \cdots, \gamma_N$, not all zero, such that for all $x \in [0, l]$,*

$$\begin{aligned}(9)\qquad &\max_{h \le \rho \le H} \{\gamma_0 F_0(x, \rho) + \gamma_1 F_1(x, \rho) + \cdots + \gamma_N F_N(x, \rho))\} \\ &\qquad = \gamma_0 F_0(x, \rho^*(x)) + \gamma_1 F_1(x, \rho^*(x)) + \cdots + \gamma_N F_N(x, \rho^*(x)).\end{aligned}$$

Conversely, if a function $\rho^*(x)$ and constants $\gamma_0 > 0$, and $\gamma_1, \gamma_2, \cdots \gamma_N$ exist satisfying (9) *and if the conditions*

$$\int_0^l F_i(x, \rho^*(x))\,dx = V_i\,, \qquad h \le \rho^*(x) \le H$$

hold, then $\rho^*(x)$ solves Problem 5.

We will use the converse part of Theorem 3 and methods similar to those used in [3],[4] to solve Problem 4. Using $\rho = (a, b)$ and selecting convenient values for the Lagrange Multipliers γ_i we find, after a bit of manipulation, that the extremal condition for (a, b) as a function of y, z can be obtained as the solution of the following elementary minimum problem.

Problem 6. For any given $y, z \ge 0$, find the values of a, b that minimizes the function $r(a, b)$ defined by

$$r(a, b) = \frac{y^2}{a^3 b} + \frac{z^2}{a b^3} + ab,$$

over all (a, b) in the rectangle $H_1 \le a \le H_2, \quad H_3 \le b \le H_4$.

The minimizing point for r is uniquely defined and is a continuous function of y, z. Thus we can express the extremal condition for Problem 4 in the general form

$$(10)\qquad a = S_1(y, z), \qquad b = S_2(y, z).$$

More precisely, the values for $S_1(y, z)$ and $S_2(y, z)$ that minimize r are selected from the following five possibilities.

If the minimum is inside the rectangle, then solving $\partial r/\partial a = 0$ and $\partial r/\partial b = 0$ simultaneously gives

$$(11)\qquad a = \sqrt[3]{2y^2/z}, \qquad b = \sqrt[3]{2z^2/y}.$$

If the minimum is on the bottom or the top of the rectangle, then either $b = H_3$ or $b = H_4$ and solving $\partial r/\partial a = 0$ gives

$$(12)\qquad a = \max\left\{H_1,\ \min\left\{H_2,\ \sqrt{z^2 + \sqrt{z^4 + 12b^6y^2}}\Big/\sqrt{2b^2}\right\}\right\}.$$

If the minimum is on the left or right edge then either $a = H_1$ or $a = H_2$ and solving $\partial r/\partial b = 0$ gives

$$b = \max\left\{H_3,\ \min\left\{H_4,\ \sqrt{y^2 + \sqrt{y^4 + 12a^6z^2}}\Big/\sqrt{2}a^2\right\}\right\}. \tag{13}$$

Now (10) can now be inserted into (4), (5) to obtain a pair of nonlinear second-order differential equations for y, z which we write in the vector form

$$\vec{U}' = \vec{G}\left(x, \vec{U}\right), \qquad \text{where} \qquad \vec{U} = (y, yp, z, zp), \qquad y' = yp, \quad z' = zp. \tag{14}$$

The solution of this system will then give $a(x)$, $b(x)$, using (10).

2.2. The numerical solutions. The equations (14) are, for the most part, numerically well behaved. In order to begin the solution of (14) we need four boundary conditions at $x = 0$. We choose two parameters y_0 and z_0' and use the given boundary conditions (4), (5) to obtain the four initial conditions

$$\vec{U}\Big]_{x=0} = (y_0, -y_0, 0, z_0').$$

The three values y_0, z_0', and λ are then used as shooting parameters to satisfy the two boundary conditions at $x = l$ and the volume condition (6). This gives a system of three equations in three unknowns of the form $\vec{F}(\vec{\alpha}) = \vec{0}$ where

$$\vec{\alpha} = (y_0, z_0', \lambda), \qquad \vec{F}(\vec{\alpha}) = \left(y(l),\ z(l),\ \widehat{V} - \int_0^l a(x)b(x)\,dx\right).$$

The numerical solution of this problem was quite straightforward. Some canned subroutines from the standard computing package CMLIB were used. The initial value problem (14) was solved using DEBRKF and the subroutine SNSQE was used to solve the three simultaneous equations $\vec{F}(\vec{\alpha}) = \vec{0}$. A simple FORTRAN program was written, which defined the equations, called up the subroutines from CMLIB and output the results. Some computer-generated plots of the extremals are given in Fig. 1 where the curves are labeled as follows: 1 is b, 2 is ab, 3 is a, 4 is z, 5 is y, 6 is $1/(ab^3)$, and 7 is $1/(a^3b)$. An artist's visualization[1] of this three-dimensional shape is also given in the left-hand side of Fig. 2. The other two columns are the optimal shapes for other boundary conditions. The middle one is when both ends of the column are supported by journal bearings where the shafts are at right angles to each other. The one on the right is when both ends are supported by journal bearings having the two shafts parallel to each other.

The only real difficulty with the method was that a very good first guess (sometimes to within 30 percent or so) had to be used for $\vec{\alpha}$ in order to get convergence from SNSQE. Depending on the accuracy of the first guess (if it produced convergence), it would generally require less than 30 to 60 seconds of CPU time on a VAX-11/750 computer to get the solution when using 100 mesh points.

SNSQE works by simply minimizing the norm of $\vec{F}$. This sometimes gave erroneous solutions for which the norm was locally minimized but was not small. There were also invalid solutions having a very small norm, but either y or z had too many zeros in the interval. Occasional difficulties arose with the calculation of y^2/z and z^2/y near $x = l$ where both vanish. In this case, it was easy to simply recognize that we should have $a(x) = H_1$ and $b(x) = H_3$.

[1] Figure 2 was drawn by Jack Snowden.

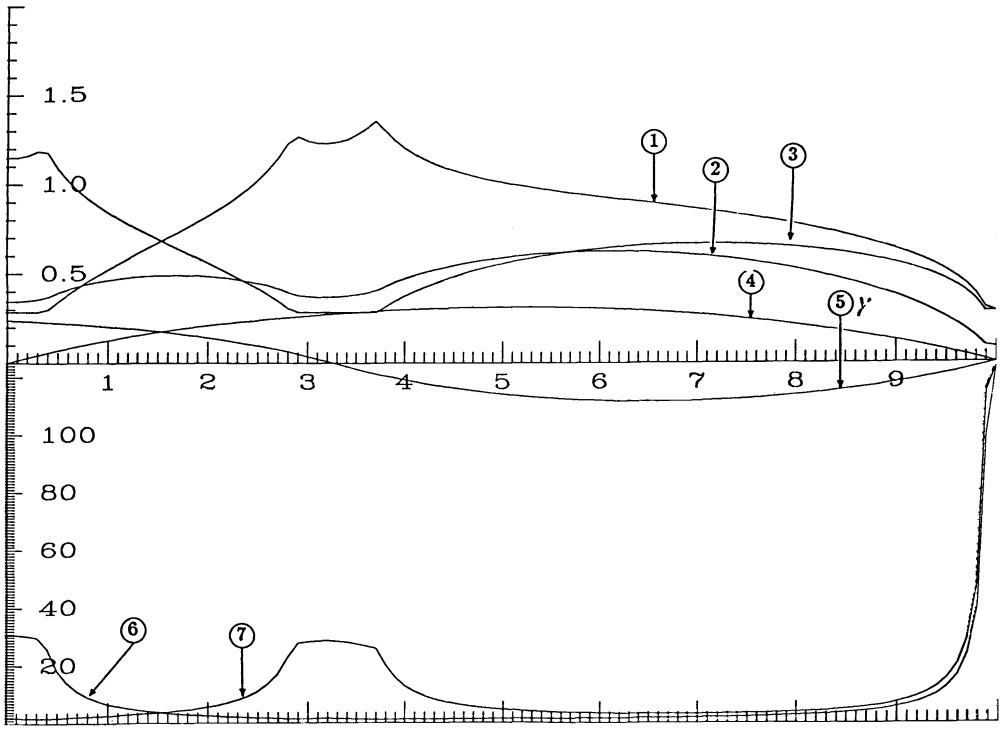


FIG. 1. *Graphs for the clamped-pinned-pinned-pinned column with* $H_1 = H_3 = .3$, $H_2 = H_4 = \infty$, $l = 1$, $y(0) = .2519$, $z'(0) = 1.596$, $\lambda = 5.023$, $\widehat{V} = .5$.

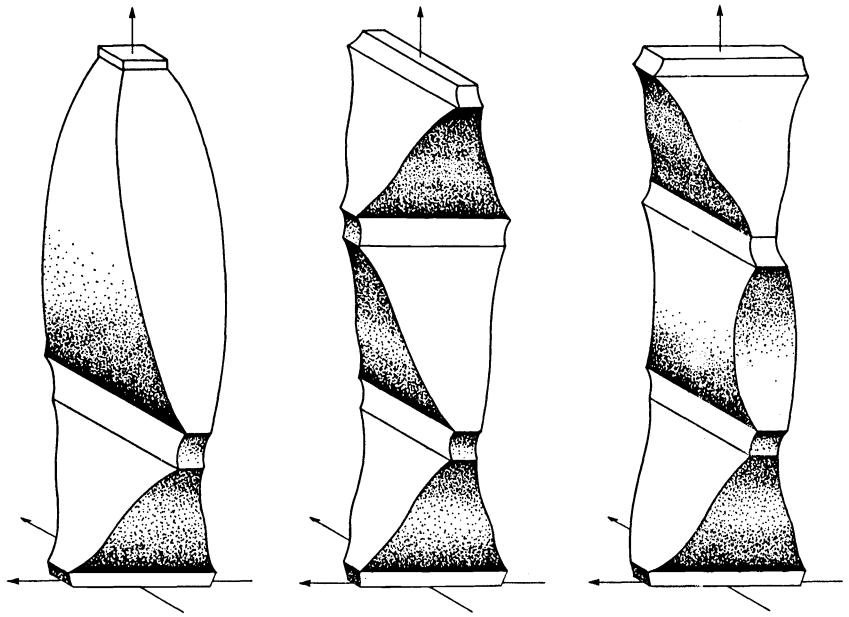


FIG. 2. *The shapes of some strongest columns using a square for* $\widehat{D}$.

2.3. Discussion of the solutions and generalizations. Many results in the literature (for example, [1],[2] and the references given there) use arguments involving rearrangements to prove certain kinds of symmetry relationships for the solution of some extremal eigenvalue problems. Although, the shape of this column (given in the left-hand side of Fig. 2) is not at all symmetric, it is interesting to note that, for most choices of H_i, the computed values for the eigenfunction z were observed to satisfy the relation $z(x) = z(l - x)$ to within a few percent. In a similar way, if x^* is the interior zero of the eigenfunction y then the symmetry relations $y(x^* - x) = -y(x^* + x)$ and $y(l - x) = y(x^* + x)$ are "almost" satisfied. Considering the strange and nonsymmetrical shape of the column on the left side of Fig. 2, it is, on the one hand, remarkable that the eigenfunctions displayed such a high degree of symmetry. On the other hand, this is the optimal shape, so a nearly symmetrical buckling mode is to be expected. In fact, it was observed that the solutions y, z became more and more symmetric as $H_1, H_3 \to 0$ and $H_2, H_4 \to \infty$. It also appeared that the graph of the cross-sectional area $a(x)b(x)$ was smooth at the points near $x = \frac{1}{3}$ where the other functions had corner points.

There are many different combinations of boundary conditions that could be imposed on the column besides those considered so far. It could be either clamped, pinned, or free in either the x-y plane or in the x-z plane at either $x = 0$ or $x = l$. The extremal condition (10) is the same in any case. The only difference occurs in the boundary conditions to be imposed on y and z. Figure 2 shows the optimal shapes for some other possible combinations of boundary conditions.

Suppose we consider the example where the column is pinned in all directions at both $x = 0$ and $x = l$. Let us also take $H_1 = H_3 = 0$ and $H_2 = H_4 = \infty$.[2] Because of the symmetric nature of the problem, the extremals will satisfy $a(x) = b(x)$ and $y = z$. The extremal condition (10) reduces to $a^3(x)b^3(x) = 4yz$, the singularities have canceled out, and, using $A(x) = \hat{A}a(x)b(x)$, we see that $A^3(x) = 4\hat{A}^3y^2$. When y is properly renormalized, this is the same extremal condition used by Keller [12].

It is interesting to note that if $H_1 > 0$ and $H_3 > 0$ are fixed, then H_2 and H_4 will not impinge on the extremal shapes if they are large enough. That is, the computed values for the extremal solution $a(x)$, $b(x)$ will automatically satisfy $a(x) < H_2$ and $b(x) < H_4$ for all x, so we can take $H_2 = H_4 = \infty$. Also, if H_2 and H_4 are fixed and if the eigenfunctions y, z do not have a common zero, then the lower bounds H_1 and H_2 will not impinge on the extremal shape if they are small enough. A proof of these statements can be based on (11)-(13). Another way to think about this is to hold all the bounds H_i fixed and change $\hat{V}$. If $\hat{V}$ is small enough, then H_2 and H_4 will not impinge. If, however, $\hat{V}$ is large enough and if, in addition, y, z do not have a common zero, then H_1 and H_3 will not impinge. If, however, y, z have a common zero, then H_1 and H_3 will always impinge. In any of these cases, at least one of the constraints will always impinge.

2.4. Buckling in other planes. We now propose to consider buckling in some plane other than the x-y or x-z planes. Suppose that $\hat{\mathcal{D}}$ is a very thin rectangle with its axis inclined at an angle of $\pi/4$ to the x-z plane. In this case the column will clearly not buckle first in either the x-y or x-z planes as expected. We must take care of this possibility by imposing some further restrictions on the geometry of $\hat{\mathcal{D}}$.

Let P_θ be a plane containing the x axis including an angle θ between P_θ and the x-z plane. Let $I(x, \theta)$ denote the moment of inertia of $\mathcal{D}_x$ about a line through its

[2] One must, however, be careful when using $H_1 = H_3 = 0$. There are unexpected problems associated with such things [6].

centroid and perpendicular to P_θ so that

$$I(x,\theta) = \iint_{D_x} r^2\,dy\,dz, \qquad r = |y\sin\theta - z\cos\theta|.$$

It follows that

$$I(x,\theta) = a^3(x)b(x)\sin^2\theta + a(x)b^3(x)\cos^2\theta - a^2(x)b^2(x)\sin 2\theta \iint_{\widehat{D}} \widehat{y}\widehat{z}\,d\widehat{y}\,d\widehat{z}.$$

In order to rule out the case of the inclined thin rectangle considered above, and others like it, we will simply assume that $\widehat{D}$ satisfies the condition, $\iint_{\widehat{D}} \widehat{y}\widehat{z}\,d\widehat{y}\,d\widehat{z} = 0$. This could be easily achieved by rotating $\widehat{D}$ in the y-z plane and then renormalizing it so that $I_y(\widehat{D}) = I_z(\widehat{D}) = 1$. It follows that the moment of $\widehat{D}$ is 1 with respect to any line and that $I(x,\theta) = a^3(x)b(x)\sin^2\theta + a(x)b^3(x)\cos^2\theta$. Now we see that if θ moves from 0 to $\pi/2$ then $I(x,\theta)$ is a monotonic function, moving between the two values $a^3(x)b(x)$ and $a(x)b^3(x)$.

Suppose the column buckles in the P_θ plane and let $w(x)$ be the displacement in the buckled state. At $x = 0$ the boundary conditions on w should be intermediate between being pinned and clamped. Such conditions are given by

$$w(0) = (1-\alpha)w''(0) + \alpha w'(0) = 0, \qquad w(l) = w''(l) = 0.$$

The number $\alpha \in [0,1]$ measures the "hardness" of the support for the column at $x = 0$ in the P_θ plane. It is a function of θ so that, $\alpha = \alpha(\theta)$. We now introduce the bending moment $v = I(x,\theta)w''$ and, using the equations given by Barnes [2], we obtain the boundary conditions and the eigenvalue problem for the determination of the buckling load in the P_θ plane

$$(15) \qquad v'' + \frac{\Lambda}{I(x,\theta)}v = 0, \qquad \alpha I(0,\theta)[v(0) + v'(0)] - \Lambda(1-\alpha)v(0) = v(l) = 0.$$

Note that the eigenvalue Λ occurs in both the boundary conditions and in the differential equation. In order to ensure that the solution of Problem 2 determined above really is maximal, we need to guarantee that the column will buckle first in the x-y and x-z planes and not in any P_θ plane. To do this we select a value for $\alpha = \alpha_0(\theta)$ so that the second eigenvalue Λ of (15) is equal to the optimal value λ^*, determined above. That is, we make the support as hard as is necessary. Then, $\alpha_0(\theta)$ will define the lower limit on α for which these solutions are really the optimal ones. It would be simple enough to calculate this value in any given example using (15).

These observations show one way to deal with buckling in the P_θ plane. Another way would be to optimize over each plane P_θ for any given support function $\alpha(\theta)$. We leave that problem for the interested reader to pursue.

REFERENCES

[1] D. C. BARNES, *Lower bounds for eigenvalues of Sturm-Liouville systems*, Indiana J. Math., 30 (1981), pp. 193-198.

[2] ———, *Buckling of columns and rearrangements of functions*, Quart. Appl. Math., 41 (1983), pp. 169-180.

[3] ———, *Extremal problems for eigenvalues with applications to buckling, vibration and sloshing*, SIAM J. Math. Anal., 16 (1985), pp. 341-357.

[4] ———, *Extremal problems for eigenvalue functionals*, SIAM J. Math. Anal., 16 (1985), pp. 1284-1294.

[5] ———, *Some approximation formula for stochastic eigenvalues*, SIAM J. Math. Anal., 18 (1987), pp. 933-940.

[6] D. C. BARNES, *The shape of the strongest column is arbitrarily close to the shape of the weakest column*, Quart. Appl. Math., to appear.
[7] E. R. BARNES, *The shape of the strongest column and some related extremal eigenvalue problems*, Quart. Appl. Math., 34 (1977), pp. 393-409.
[8] T. CLAUSEN, *Über die form architektonisher saulen*, Bulletin Physico-Mathematiques et Astronumiques Tome 1, (1849-1853), pp. 279-294.
[9] R. D. GENTRY AND D. O. BANKS, *Bounds for functions of eigenvalues*, J. Math. Anal. Appl., 51 (1975), pp. 100-128.
[10] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[11] K. HOLMÅKER, *Some mathematical aspects on a problem of the optimal design of a vibrating beam*, SIAM J. Math. Anal., 18 (1987), pp. 1367-1377.
[12] J. B. KELLER, *The shape of the strongest column*, Arch. Rat. Mech. Anal., 5 (1960), pp. 275-285.
[13] ———, *The minimum ratio of two eigenvalues*, SIAM J. Appl. Math., 31 (1976), pp. 485-491.
[14] J. B. KELLER AND F. I. NIORDSON, *The tallest column*, J. Math. Mech., 16 (1966), pp. 433-446.
[15] M. G. KREIN, *On certain problems on the maximum and minimum of characteristic values and on Lyapunov zones of stability*, Trans. Amer. Math. Soc., 2 (1955), pp. 163-187.
[16] J. L. DE LAGRANGE, *Sur la figure des colonnes*, Miscellanea Tourinesia, (Royal Society of Turin), Tomus V, 1770-1773, p. 123; also *Oevures*, 2, 1770-1773, pp. 125-170.
[17] M. K. MYERS AND W. R. SPILLERS, *A note on the strongest fixed-fixed column*, Quart. Appl. Math., 44 (1986), pp. 583-588.
[18] I. TADJBAKHSH AND J. B. KELLER, *Strongest columns and isoperimetric inequalities for eigenvalues*, J. Appl. Mech., 29 (1962), pp. 159-164.
[19] S. P. TIMOSHENKO AND J. N. GERE, *Theory of Elastic Stability*, McGraw-Hill, New York, 1961.
[20] J. L. TROUTMAN, *Variational Calculus With Elementary Convexity*, Springer-Verlag, New York, 1983.
[21] B. E. WILLNER AND T. J. MAHOR, *The two-dimensional eigenvalue range and extremal eigenvalue problems*, SIAM J. Math. Anal., 13 (1982), pp. 621-631.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

013

CANONICAL FACTORIZATIONS OF DISCONJUGATE DIFFERENTIAL OPERATORS—PART II*

ANTONIO GRANATA†

Abstract. This paper, which completes the discussion in Part I [*SIAM J. Math. Anal.*, 11 (1980), pp. 160–172], provides the following: (1) some characterizations of those operators that are disconjugate on an open interval $\mathcal{T}$ and that admit only one Pólya–Mammana factorization on $\mathcal{T}$ (constant factors apart); (2) a sufficient condition for the existence, on a given interval of disconjugacy, of a double canonical factorization of type (II) together with some counterexamples showing that the given condition is likely to be necessary. A detailed nontrivial treatment of factorizations of second-order operators will guide and highlight the whole discussion.

Key words. disconjugacy, canonical factorizations

AMS(MOS) subject classifications. primary 34C10; secondary 34C99

Introduction. This paper is a continuation of a previous work [3] and solves some of the problems left open therein. The numbering of the sections and formulas follows that given in [3]. As in the first part, operators of type (*) are considered; see formula (1.1) and subsequent lines in [3]. The particular factorizations discussed are always of type (1.3)–(1.4); it is tacitly assumed that these factorizations are *global* on the specified interval: [3, p. 162, line 5]. Disconjugacy on a *closed* interval $[a, b]$, $-\infty \leqq a < b \leqq +\infty$ (closed in the topology of the extended real line), is meant in the sense given by Definition 3.5. Besides the symbol $D_n(a, b)$ (see Definition 2.1), we use the symbols $D_n[a, b)$, $D_n(a, b]$, and $D_n[a, b]$, all of whose meanings are obvious. It must be kept in mind that the three relationships $L \in D_n(a, b)$, $L \in D_n[a, b)$, and $L \in D_n(a, b]$ are equivalent to each other [6, Lemma 2.3]. On the other hand, $L \in D_n[a, b]$ is in general a stronger property than $L \in D_n(a, b)$; for example, the operator d^n/dt^n belongs to $D_n(-\infty, +\infty)$ but not to $D_n[-\infty, +\infty]$. The following definition can be useful.

DEFINITION. The symbol $\tilde{D}_n[a, b]$ denotes the set of all the operators $L \in D_n(a, b)$ such that L has the properties stated in Theorem 3.3.

The locutions "one" or "essentially unique" (in quotes) are used to mean "one apart from constant factors" when referring to a function, a set of functions, or the coefficients p_i of a factorization (1.3). C.F. is used instead of *canonical factorization* for the sake of brevity.

6. Factorizations of second-order operators. We discuss some nonobvious results concerning second-order operators that clarify the main results in the following sections.

A. A description of all the possible factorizations. Given a linear second-order operator

$$L_2 u \equiv u'' + a_1(t)u + a_0(t)u; \quad a_i \in L^1_{\text{loc}}(a, b) \quad (i = 1, 2), \tag{6.1}$$

we shall describe all its possible factorizations on (a, b) of type

$$L_2 u \equiv p_2[p_1(p_0 u)']' \quad \forall u \in AC^1(a, b) \tag{6.2}$$

for some suitable strictly positive functions $p_i(t)$, $i = 1, 2, 3$.

* Received by the editors December 8, 1986; accepted for publication (in revised form) October 10, 1987. This paper was written while the author was a member of the Italian Committee for Scientific Research C.N.R.-G.N.A.F.A.

† Dipartimento di Matematica, Università della Calabria, 87036 Rende (Cosenza), Italy.

THEOREM 6.1. *If $L_2 \in D_2(a, b)$, then for each solution u_0 to $L_2u = 0$ strictly positive on (a, b), there exists a "unique" factorization of type* (6.2) *such that $p_0(t) \equiv 1/u_0(t)$. The coefficient p_1 in this factorization is given by*

$$p_1(t) = \frac{1}{p_0^2(t)} \exp\left(\int_{t_0}^{t} a_1(\tau)\, d\tau\right) \quad \text{with } t_0 \text{ fixed in } (a, b), \tag{6.3}$$

while p_2 is obviously given by $p_2(t) = 1/p_0p_1$. In particular all the factorizations of the operator $u'' + q(t)u$, $q \in L^1_{\text{loc}}(a, b)$, are given by

$$u'' + q(t)u \equiv \frac{1}{u_0(t)}\left[u_0^2(t)\left(\frac{u}{u_0(t)}\right)'\right]', \qquad u_0(t) > 0 \quad \text{on } (a, b). \tag{6.4}$$

Proof. Our simple argument completes the standard one used to show the existence of a factorization (6.2) (cf. [1, p. 6] or [7, p. 316]). If (6.2) holds true, the function $u_0(t) \equiv 1/p_0(t)$ must be a solution strictly positive on (a, b), whence it follows that all the possible factorizations can be obtained by considering all the possible choices for p_0. When p_0 has been fixed, we infer from the identity (6.2) that

$$L_2u = 0 \Leftrightarrow p_2[p_1(p_0u)']' = 0$$

if and only if

$$p_1(t)[p_0(t)u(t)]' = \text{constant} \quad \forall u \text{ such that } L_2u = 0, \tag{6.5}$$

where the constant obviously depends on the choice of u. To derive p_1 from (6.5) let us suppose that u is any solution independent of $1/p_0$ (otherwise $u(t) \equiv c/p_0(t)$ and $(p_0(t)u(t))' \equiv 0$). In this case Liouville's formula gives

$$W\left(u(t), \frac{1}{p_0(t)}\right) = A \exp\left(-\int_{t_0}^{t} a_1(\tau)\, d\tau\right)$$

where $A = \text{constant} \neq 0$ and t, $t_0 \in (a, b)$. On the other hand, direct calculations give

$$W\left(u(t), \frac{1}{p_0(t)}\right) = -\frac{(p_0u)'}{p_0^2}$$

whence

$$(p_0u)' - Ap_0(t) \exp\left(-\int_{t_0}^{t} a_1(\tau)\, d\tau\right). \tag{6.6}$$

From (6.5), (6.6) it can be inferred that, up to a constant factor, p_1 is given by (6.3). □

B. Examples: Constant coefficient operators. In order to illustrate the theorem we consider the operator

$$L_2u \equiv u'' + Au' + Bu \quad \text{where } A, B \text{ are real constants.} \tag{6.7}$$

Let r_1 and r_2 be the complex roots of the characteristic equation $r^2 + Ar + B = 0$. When r_1 and r_2 are not real, then L_2 is only disconjugate on sufficiently small intervals of $\mathbb{R}$, while when both r_1 and r_2 are real then L_2 is disconjugate on $(-\infty, +\infty)$ and hence on any interval of $\mathbb{R}$. Let us limit our investigation to those cases where the roots are real.

First case. $A^2 - 4B = 0$. The operator is

$$L_2u \equiv u'' + Au' + \frac{A^2}{4}u. \tag{6.8}$$

(i) If we choose $(a, b) = (-\infty, +\infty)$, then L_2 has only "one" factorization, namely

$$L_2 u \equiv e^{-(A/2)t}[(e^{(A/2)t})']' \tag{6.9}$$

which, by Theorem 2.1, is necessarily a double C.F. of type (I).

(ii) If we choose $(a, b) \neq \mathbb{R}$ then, besides (6.9), L_2 admits of an infinite number of factorizations which are given by

$$L_2 u \equiv \frac{e^{-(A/2)t}}{\alpha + \beta t}\left[(\alpha + \beta t)^2\left(\frac{e^{(A/2)t}u}{\alpha + \beta t}\right)'\right]' \tag{6.10}$$

where α, β are such that $\alpha + \beta t > 0$ on (a, b). Hence it can be seen that on any fixed interval (a, b), L_2 has an infinite number of essentially different double C.F.s of type (II).

Second case. $A^2 - 4B > 0$. Assuming that $r_1 < r_2$, we have the following three types of factorizations:

$$L_2 u \equiv e^{r_1 t}[e^{(r_2 - r_1)t}(e^{-r_2 t}u)']' \quad \text{on } (-\infty, +\infty), \tag{6.11}$$

which is of type (I) at $-\infty$ and of type (II) at $+\infty$;

$$L_2 u \equiv e^{r_2 t}[e^{(r_1 - r_2)t}(e^{-r_1 t}u)']' \quad \text{on } (-\infty, +\infty), \tag{6.12}$$

which is of type (II) at $-\infty$ and of type (I) at $+\infty$;

$$L_2 u \equiv \frac{e^{(r_1+r_2)t}}{\alpha e^{r_1 t} + \beta e^{r_2 t}}\left[\frac{(\alpha e^{r_1 t} + \beta e^{r_2 t})^2}{e^{(r_1+r_2)t}}\left(\frac{u}{\alpha e^{r_1 t} + \beta e^{r_2 t}}\right)'\right]' \tag{6.13}$$

where α, β are such that $\alpha e^{r_1 t} + \beta e^{r_2 t} > 0$ on the chosen interval (a, b). Hence, corresponding to all the possible couples α, $\beta > 0$, we obtain an infinite number of factorizations valid on any interval (a, b), all of which are double C.F.s of type (II) on (a, b), including the case $(a, b) = (-\infty, +\infty)$.

C. Theoretical results. The above examples provide a good illustration of the following general results.

THEOREM 6.2. *Let* $L_2 \in D_2(a, b)$.

(I) *The following properties are equivalent*:

(i) $L_2 \in \tilde{D}_2[a, b]$;

(ii) $L_2 u = 0$ *has only "one" solution* u_0 *that is strictly positive on* (a, b);

(iii) L_2 *has only "one" factorization on* (a, b) *of type* (1.3)-(1.4), *which is necessarily a double* C.F. *of type* (I).

(II) L_2 *is disconjugate on* $[a, b]$ *if and only if it has infinitely many essentially different double* C.F.*s of type* (II).

In the case where $L_2 \in D_2[a, b]$, we can give a complete description of the various types of factorizations by relating them to a mixed hierarchical system of solutions.

THEOREM 6.3. *Let* $L_2 \in D_2[a, b]$ *and let* $(\bar{u}_1, \bar{u}_2)$ *be the mixed hierarchical system of solutions to* $L_2 u = 0$ *such that*

$$\begin{aligned} &\bar{u}_1 \ll \bar{u}_2, \qquad t \to a, \\ &\bar{u}_2 \ll \bar{u}_1, \qquad t \to b. \end{aligned} \tag{6.14}$$

Then:

(I) L_2 *has a* ("*unique*") *factorization on* (a, b) *of form*

$$L_2 u = p_2\left[p_1\left(\frac{u}{\bar{u}_1}\right)'\right]', \tag{6.15}$$

which is the C.F. *of type* (I) *at a.*

Similarly L_2 has a ("unique") factorization on (a, b) of form

$$L_2u \equiv q_2\left(q_1\left(\frac{u}{\bar{u}_2}\right)'\right]', \tag{6.16}$$

which is the C.F. *of type* (I) *at b.*

(II) *If $\bar{u}$ is any solution to $L_2u = 0$ of type*

$$\bar{u}(t) = \alpha\bar{u}_1(t) + \beta\bar{u}_2(t) \quad (\alpha, \beta = \textit{positive constants}), \tag{6.17}$$

then L_2 has a ("unique") factorization on (a, b)

$$L_2u \equiv r_2\left(r_1\left(\frac{u}{\bar{u}}\right)'\right]', \tag{6.18}$$

which is a double C.F. *of type* (II).

Proofs of Theorems 6.2 *and* 6.3. It is known that the operator L_2 is disconjugate on (a, b) if and only if $L_2u = 0$ has a solution u_0 that is strictly positive on (a, b). Two contingencies can arise: there exists only "one" strictly positive solution on (a, b) or there are two linearly independent such solutions (hence infinitely many, essentially different such solutions). Now the claims in Theorem 6.2 directly follow from Theorem 6.1 by considering that, for a second-order operator, each factorization is a C.F. at each separate endpoint. In order to prove Theorem 6.3, note that the two functions $\bar{u}_1$, $\bar{u}_2$ are strictly positive on (a, b) [3, Lemma 5.5]; hence, by Theorem 6.1, L_2 has "unique" factorizations on (a, b) of form (6.15), (6.16), and (6.18). Let us examine (6.15), which implies that $L_2u = 0$ has a solution $v(t) \equiv \bar{u}_1(t)\int_{t_0}^{t} 1/p_1$, with t_0 arbitrarily fixed on (a, b). As v and $\bar{u}_1$ are linearly independent, the first relation (6.14) implies that $\bar{u}_1 = o(v)$, $t \to a$. This in turn implies that

$$\lim_{t\to a}\left|\int_{t_0}^{t}\frac{1}{p_1}\right| = +\infty, \quad \text{i.e.,} \quad \int_a \frac{1}{p_1} = +\infty.$$

The claim about (6.16) is similarly proved. Factorization (6.18) implies that $L_2u = 0$ has a solution $w(t) \equiv \bar{u}(t)\int_{t_0}^{t} 1/r_1$. On the other hand, both (6.14) and (6.17) imply that $\bar{u}$ is a solution with maximal order of growth at both a and b; this means that for any solution u to $L_2u = 0$ both $\lim_{t\to a} u(t)/\bar{u}(t)$ and $\lim_{t\to b} u(t)/\bar{u}(t)$ exist as finite numbers ($=0$ or $\neq 0$). Hence we infer that both limits

$$\lim_{\substack{t\to a\\ [t\to b]}}\frac{w(t)}{\bar{u}(t)} \equiv \lim_{\substack{t\to a\\ [t\to b]}}\int_{t_0}^{t}\frac{1}{r_1}$$

exist in $\mathbb{R}$. This means that $\int_a^b 1/r_1 < +\infty$. □

In the subsequent sections we investigate to what extent the contingencies described in Theorems 6.2 and 6.3 hold for nth order operators. In § 7 we show that Theorem 6.2(I) directly extends to nth order operators. In § 8 a sufficient condition for the existence of a double C.F. of type (II) is given. Section 9 shows that the results in Theorem 6.3, when literally extended to the nth order case, are completely false. For the sake of completeness we now state a refinement of Theorem 6.2 which is interesting in the context of this section but admits of no extension to the nth order case.

PROPOSITION 6.4. *Let the operator L_2, defined by* (6.1), *be disconjugate on (a, b) and define*

$$A_1(t) \equiv \exp\left(\int_{t_0}^{t} a_1(\tau)\, d\tau\right) \tag{6.19}$$

for some fixed $t_0 \in (a, b)$, the particular choice of t_0 being immaterial. Then to the properties listed in Theorem 6.2(I), *the following can be added*:

(iv) $L_2u = 0$ *has a solution u_0 with no zeros both in a neighborhood of a and in a neighborhood of b and such that*

$$\int_a \frac{dt}{A_1(t)u_0^2(t)} = \int^b \frac{dt}{A_1(t)u_0^2(t)} = +\infty. \tag{6.20}$$

If this is the case then $L_2u = 0$ has a "unique" solution u_0 satisfying (6.20); *it coincides with "the" solution u_0 strictly positive on (a, b) and is such that*

$$u_0 \ll u_1, \qquad \begin{cases} t \to a, \\ t \to b, \end{cases} \tag{6.21}$$

for any solution u_1 linearly independent from u_0. The "only" global factorization of L_2 on (a, b) is

$$L_2u \equiv \frac{1}{A_1(t)u_0(t)}\left[A_1(t)u_0^2(t)\left(\frac{u}{u_0^2(t)}\right)'\right]' \quad \forall u \in AC^1(a, b). \tag{6.22}$$

The easy proof can be based on Theorems 3.3 and 6.2 and is left to the reader. Some explanations on the statement are needed. It is a classical fact that if L_2 is nonoscillatory at a $[b]$—which is the case if and only if it is disconjugate on a one-sided neighborhood of a $[b]$—then there is a "unique" solution u_a $[u_b] \neq 0$ in a neighborhood of a $[b]$ such that

$$\int_a \frac{dt}{A_1(t)u_a^2(t)} = +\infty, \qquad \left[\int^b \frac{dt}{A_1(t)u_b^2(t)} = +\infty\right] \tag{6.23}$$

(see Hartman [4_{bis}, p. 355]). If u_a and u_b both exist it can happen that $u_a \neq u_b$ even if $L_2 \in D_2(a, b)$. Proposition 6.4 provides some characterizations for the contingency $u_a = u_b$ to occur and completes the results in Hartman [4_{bis}, Thm. 6.4, p. 355].

7. Operators that admit of only "one" factorization. This section proves the following conjecture made in [3, § 3, p. 165]: the operators described in Theorem 3.3 are those that admit of only "one" global Pólya-Mammana factorization.

THEOREM 7.1. *If $L \in D_n(a, b)$ then to the properties listed in Theorem* 3.3 *the following can be added*:

$(2)_{\text{bis}}$. *L has only "one" factorization of type* (1.3)-(1.4) *on* (a, b).

$(4)_{\text{bis}}$. *If u, v are any two nontrivial solutions to $Lu = 0$ such that $u = o(v)$, $t \to a$, then the relation $u = o(v)$, $t \to b$, also holds true. By interchanging the roles of a and b a similar result is obtained.*

$(5)_{\text{bis}}$. *If u, v are any two nontrivial solutions to $Lu = 0$ such that $u \sim c_1v$, $t \to a$, (for a suitable constant $c_1 \neq 0$), then the relation $u \sim c_2v$, $t \to b$, also holds true (for another suitable constant $c_2 \neq 0$).*

The roles of a and b can be interchanged.

Proof. The inferences $(4)_{\text{bis}} \Rightarrow (4)$ and $(5)_{\text{bis}} \Rightarrow (5)$ are obvious.

$(4) \Rightarrow (4)_{\text{bis}}$. Let $(u_1, \cdots, u_n)$ be some fixed double hierarchical system to $Lu = 0$ on (a, b), i.e.,

$$u_1 \ll u_2 \ll \cdots \ll u_n \quad \text{both as } t \to a \text{ and as } t \to b. \tag{7.1}$$

The hypothesis on u, v implies that there exist two indices j, $k \in \{1, \cdots, n\}$, $j < k$, such that

$$\begin{aligned} u &\sim c_ju_j, \qquad t \to a, \\ v &\sim c_ku_k, \qquad t \to b, \end{aligned} \tag{7.2}$$

where c_j, c_k are suitable constants. From (7.1) and (7.2) we infer that $u \ll v$ both as $t \to a$ and as $t \to b$.

(5)$\Rightarrow$(5)$_{\text{bis}}$. Let $(u_1, \cdots, u_n)$ and $(v_1, \cdots, v_n)$ be two fixed double hierarchical systems; hence

$$\begin{aligned} u_i &\sim \alpha_i v_i, \qquad t \to a, \\ u_i &\sim \beta_i v_i, \qquad t \to b, \end{aligned} \tag{7.3}$$

where α_i, β_i are nonzero constants. On the other hand, for a suitable couple of indices j, $k \in \{1, \cdots, n\}$ we have the relations

$$u \sim \alpha u_j, \quad t \to a, \qquad v \sim \beta v_k, \quad t \to a \quad (\alpha, \beta \neq 0). \tag{7.4}$$

The hypothesis $u \sim c_1 v$, $t \to a$, implies that $j = k$, and from (7.3) and (7.4) we infer that $u \sim c_2 v$, $t \to b$.

(2)$_{\text{bis}}\Rightarrow$(2). It is obvious from Theorem 2.1.

(2)$\Rightarrow$(2)$_{\text{bis}}$. This is true for $n = 2$ (see Theorem 6.2). Assume that $n \geqq 3$ and that

$$Lu = p_n[p_{n-1}(\cdots(p_0 u)' \cdots)']' \tag{7.5}$$

is any global factorization of L on (a, b). We shall prove that (7.5) is necessarily a double C.F. of type (I) on (a, b). As a first step we show that the contingency cannot arise: $\int_a 1/p_1 < +\infty$. If it did arise the two solutions

$$u_0(t) = \frac{1}{p_0(t)} \quad \text{and} \quad u_1(t) = \frac{1}{p_0(t)} \int_a^t \frac{1}{p_1}$$

would verify the relation $u_1 \ll u_0$ as $t \to a$ and also as $t \to b$ (by property (4)$_{\text{bis}}$). But this is impossible since

$$\lim_{t \to b} \frac{u_1(t)}{u_0(t)} = \lim_{t \to b} \int_a^t \frac{1}{p_1} = \int_a^b \frac{1}{p_1} > 0 \quad (\text{possibly} = +\infty).$$

It thus follows that $\int_a 1/p_1 = +\infty$.

Now we want to show that $\int_a 1/p_i = +\infty$ $(i = 1, \cdots, n-1)$. If this were not the case, there would exist a number $k \in \{1, \cdots, n-2\}$ such that

$$\int_a \frac{1}{p_i} = +\infty, \quad i = 1, \cdots, k, \qquad \int_a \frac{1}{p_{k+1}} < +\infty. \tag{7.6}$$

Let us consider the $(k+2)$ solutions

$$u_0(t) = \frac{1}{p_0(t)}, \quad u_i(t) = \frac{1}{p_0(t)} \int_\alpha^t \frac{1}{p_1} \cdots \int_\alpha^{t_{k-1}} \frac{1}{p_k}, \quad i = 1, \cdots, k,$$

$$u_{k+1}(t) = \frac{1}{p_0(t)} \int_\alpha^t \frac{1}{p_1} \cdots \int_\alpha^{t_{k-1}} \frac{1}{p_k} \int_a^{t_k} \frac{1}{p_{k+1}}$$

for some fixed α, $a < \alpha < b$. From (7.6) we infer that

$$u_0 \ll u_1 \ll \cdots \ll u_k, \qquad t \to a \tag{7.7}$$

and also as $t \to b$, by property (4)$_{\text{bis}}$. From (7.7) we infer at once that

$$\int_\alpha^b \frac{1}{p_1} = \lim_{t \to b} \frac{u_1(t)}{u_0(t)} = +\infty.$$

Hence it is legitimate to use l'Hôpital's rule in evaluating the following limit:

$$\lim_{t\to b}\frac{u_2(t)}{u_1(t)}\equiv\lim_{t\to b}\int_\alpha^t\frac{1}{p_1}\int_\alpha^{t_1}\frac{1}{p_2}\Big/\int_\alpha^t\frac{1}{p_1}=\lim_{t\to b}\int_\alpha^t\frac{1}{p_2}.$$

As the limit on the left-hand side is $+\infty$, by (7.7), we derive $\int_\alpha^b 1/p_2=+\infty$.

By iterating the procedure we can show that

$$\int^b\frac{1}{p_i}=+\infty,\qquad i=1,\cdots,k. \tag{7.8}$$

As a conclusive step we investigate the asymptotic behavior, as $t\to b$, of the ratio u_{k+1}/u_k by two different methods giving rise to a contradiction.

First method. Relations (7.8) allow us to iterate l'Hôpital's rule in evaluating the following limit:

$$\begin{aligned}\lim_{t\to b}\frac{u_{k+1}(t)}{u_k(t)}&\equiv\lim_{t\to b}\int_\alpha^t\frac{1}{p_1}\cdots\int_\alpha^{t_{k-1}}\frac{1}{p_k}\int_a^{t_k}\frac{1}{p_{k+1}}\Big/\int_\alpha^t\frac{1}{p_1}\cdots\int_\alpha^{t_{k-1}}\frac{1}{p_k}\\ &=\lim_{t\to b}\int_\alpha^t\frac{1}{p_2}\cdots\int_\alpha^{t_{k-1}}\frac{1}{p_k}\int_a^{t_k}\frac{1}{p_{k+1}}\Big/\int_\alpha^t\frac{1}{p_2}\cdots\int_\alpha^{t_{k-1}}\frac{1}{p_k}\\ &=\cdots=\lim_{t\to b}\int_a^t\frac{1}{p_{k+1}}\equiv A>0\quad(\text{possibly } A=+\infty).\end{aligned} \tag{7.9}$$

Second method. By a similar procedure, using (7.6), we can show that

$$\lim_{t\to a}\frac{u_{k+1}(t)}{u_k(t)}=\lim_{t\to a}\int_a^t\frac{1}{p_{k+1}}=0,$$

and hence, by property $(4)_{\text{bis}}$, that $\lim_{t\to b}u_{k+1}(t)/u_k(t)=0$. This contradicts (7.9) and shows that the contingency (7.6) cannot arise; thus the proof is complete. □

Remarks. (1) Properties $(4)_{\text{bis}}$, $(5)_{\text{bis}}$ mean that the asymptotic behavior of a solution with respect to another solution is the same at both endpoints. (2) Property $(2)_{\text{bis}}$ has a certain historical interest. From some of the early work on factorizations, Frobenius [2], Ince [5, p. 125], and others, we get the distinct impression that these authors unconsciously believed that Pólya-Mammana factorizations were not only valid for any operator (which is false; see Theorem 1.1), but also "unique" (which is again false). This naive conviction led Ince to an inconclusive proof. We refer the interested reader to the introduction in [4], which provides a more detailed historical analysis.

Example. If L is the constant coefficient operator (4.1), considered on the whole real line, then $L\in\tilde{D}_n[-\infty,+\infty]$ if and only if its characteristic equation has a root of multiplicity n, say λ. If this is the case then the only "one" global factorization of L on $(-\infty,+\infty)$ is given by

$$Lu\equiv e^{\lambda t}(e^{-\lambda t}u)^{(n)}. \tag{7.10}$$

8. Double canonical factorizations of type (II). To extend the result in Theorem 6.2(II) to nth order operators the most natural approach would be to ascertain the existence of infinitely many double C.F.s of type (II) by proving the natural extension of Theorem 6.3 to nth order operators. But, as shown by the simple operator d^3/dt^3 (see § 9), all such natural extensions are completely false! Hence the legitimate suspicion arises that not every operator $L\in D_n[a,b]$, $n>2$, has double C.F.s of type (II). On the other hand, there exists one particular case when this contingency is trivially true:

this happens when L is disconjugate on a larger interval than $[a, b]$, say $(a-\varepsilon, b+\varepsilon)$, $\varepsilon > 0$. In this case each factorization of L of type (1.3)-(1.4) on $(a-\varepsilon, b+\varepsilon)$ is obviously a double C.F. of type (II) on (a, b) as the coefficients p_i are continuous and strictly positive on the compact interval $[a, b]$; hence $\int_a^b 1/p_i < +\infty$, $i = 0, 1, \cdots, n$. In this section we give a less trivial sufficient condition for the existence of at least one double C.F. of type (II) by imposing a restriction on the nature of one endpoint. We have as yet no counterexample for the general case and the question still arises whether any operator $L \in D_n[a, b]$, $n > 2$, has a double C.F. of type (II), or even an infinity of them.

A. Extension of disconjugate operators. We give some preliminary results concerning the extension of a disconjugate operator to a larger interval.

LEMMA 8.1. *Let L be an nth order operator of type (*) on an open interval $(a, b) \neq \mathbb{R}$. If L is disconjugate on an interval $[\alpha, \beta] \subsetneqq [a, b]$, then there exists a positive number ε such that L is disconjugate on $(\alpha-\varepsilon, \beta+\varepsilon) \subset (a, b)$. (Agreement: $\alpha-\varepsilon = \alpha$ if $\alpha = a$; $\beta+\varepsilon = b$ if $\beta = b$.)*

Proof. Coppel [1, Lemma 7, p. 93], outlines a proof when L has continuous coefficients. In its full generality Lemma 8.1 follows easily from nontrivial results by Levin [6, § 3]. First, let us suppose that $\beta < b$ and prove that there exists $\varepsilon_1 > 0$ such that $L \in D_n[\alpha, \beta+\varepsilon_1]$. Suppose, if possible, that $L \notin D_n[\alpha, \beta+\varepsilon]$ for each $\varepsilon > 0$; as $L \in D_n[\alpha, \beta]$ by assumption, it follows that $\bar{\alpha} = \beta$, where $\bar{\alpha}$ denotes the point conjugate to α on the right (see the definition in [6, p. 70]). Now Lemma 3.3 [6, p. 71] implies that $L \notin D_n[\alpha, \beta]$, which contradicts our assumption. Extensions to the left of the endpoint a and beyond both endpoints are similarly proved. □

DEFINITION 8.1. For a given operator L of type (*) on some interval (a, b) the endpoint a [b] is called a nonsingular endpoint if $a > -\infty$ [$b < +\infty$] and the coefficients $a_i(t)$, $i = 1, \cdots, n$, in the representation (1.1) are such that

$$a_i \in L^1(a, a+\varepsilon) \quad [a_i \in L^1(b-\varepsilon, b)] \tag{8.1}$$

for some $\varepsilon > 0$. Otherwise the endpoint is termed singular.

As an immediate corollary of Lemma 8.1 we obtain the following.

LEMMA 8.2. *Suppose that $L \in D_n(a, b)$ and that b is a nonsingular endpoint for L. Then, in whatever manner we extend the coefficients $a_i(t)$, $i = 1, \cdots, n$, of the representation (1.1) to the interval $(a, +\infty)$ by means of functions $\tilde{a}_i \in L^1_{\mathrm{loc}}(a, +\infty)$, there exists a number $\varepsilon > 0$ such that the operator $\tilde{L}$, defined by (1.1), where the coefficients a_i's are replaced by $\tilde{a}_i$'s, is disconjugate on $(a, a+\varepsilon)$. (The number ε will of course depend on the particular extension chosen.)*

Similar versions of Lemma 8.2 are to be found for cases where a is nonsingular or both a and b are nonsingular.

B. Double C.F.s of type (II). Where second-order operators are concerned the property of having infinitely many double C.F.s of type (II) has no relationship to the nature of the two endpoints. This property holds when both endpoints are nonsingular or when only one is nonsingular (see the factorizations of the operator d^2/dt^2 on any interval $\mathcal{I} \neq \mathbb{R}$ [3, p. 162]), or when both endpoints are singular as shown by

$$u'' + \frac{2}{t}u' \equiv \frac{1}{\alpha t^2 + \beta t}\left[(\alpha t+\beta)^2\left(\frac{u}{\alpha+\beta/t}\right)'\right]' \qquad (\alpha, \beta > 0) \tag{8.2}$$

on the interval $(0, +\infty)$.

For nth order operators, $n > 2$, we shall show the existence of a double C.F. of type (II) when the nature of the endpoints is restricted; the result has been stated without proof in [3, p. 164].

THEOREM 8.3. *If $L \in D_n(a, b)$ and if at least one endpoint is nonsingular then to the properties listed in Theorems* 3.1 *and* 3.2 *the following can be added*:

$$L \text{ has a double C.F. of type (II) on } (a, b). \tag{8.3}$$

Proof. Let us suppose for the sake of argument that b is a nonsingular endpoint. We have only to prove that if $L \in D_n[a, b]$ then (8.3) holds. By Lemma 8.2 there exists an extension of L, say $\tilde{L}$, which is disconjugate on some interval $(a, \tilde{b})$, $\tilde{b} > b$. Let t_0 be any number such that $b < t_0 < \tilde{b}$. As $L \in D_n[a, t_0]$, Theorem 3.1 implies that $\tilde{L}$ has a C.F. on (a, t_0) of type (II) at a. On the other hand, the coefficients of such a factorization are continuous and strictly positive on (a, t_0). Hence if we consider this factorization on the interval (a, b) only, we see that it is a double C.F. of L of type (II). □

9. Counterexamples. The following proposition shows that a direct literal extension of Theorem 6.3 to nth order operators fails to hold true.

PROPOSITION 9.1. *Let $L \in D_n[a, b]$ and let $(\bar{u}_1, \cdots, \bar{u}_n)$ be the mixed hierarchical system*

$$\begin{aligned} &\bar{u}_1 \ll \cdots \ll \bar{u}_n, && t \to a, \\ &\bar{u}_n \ll \cdots \ll \bar{u}_1, && t \to b. \end{aligned} \tag{9.1}$$

Then:

(1) *L admits of global factorizations in the following forms*:

$$Lu \equiv p_n\left[p_{n-1}\left(\cdots\left(p_1\left(\frac{u}{\bar{u}_1}\right)'\right)'\cdots\right)'\right]', \tag{9.2}$$

$$Lu \equiv q_n\left[q_{n-1}\left(\cdots\left(q_1\left(\frac{u}{\bar{u}_n}\right)'\right)'\cdots\right)'\right]'. \tag{9.3}$$

(2) *Factorizations of type* (9.2) *or* (9.3) *are in general not "unique" (they can even be infinitely many) and they are not necessarily all* C.F.*s at some endpoint.*

(3) *If $\bar{u}(t) \equiv \alpha\bar{u}_1(t) + \beta\bar{u}_n(t)$ $(\alpha, \beta > 0)$, then L admits of global factorizations in the form*

$$Lu \equiv r_n\left[r_{n-1}\left(\cdots\left(r_1\left(\frac{u}{\bar{u}}\right)'\right)'\cdots\right)'\right]'; \tag{9.4}$$

but it can happen that not even a single factorization (9.4) *is a* C.F. *at some endpoint.*

Proof. By Lemma 5.5 we may suppose that the functions $\bar{u}_1, \cdots, \bar{u}_n$ are strictly positive on (a, b). The second relation (9.1) then implies that the ordered n-tuple $(u_1, \cdots, u_n) \equiv (\bar{u}_n, \cdots, \bar{u}_1)$ satisfies condition (1.2) on (a, b); see [6, Thm. 2.1]. From this we infer that for some suitable choice of the constants $\varepsilon_i = \pm 1$, the ordered n-tuple $(v_1, \cdots, v_n) \equiv (\bar{u}_1, \varepsilon_2\bar{u}_2, \cdots, \varepsilon_n\bar{u}_n)$ also satisfies (1.2) on (a, b). This in turn implies that the ordered n-tuple $(w_1, \cdots, w_n) \equiv (\bar{u}_1, \varepsilon_2\bar{u}_2, \cdots, \varepsilon_{n-1}\bar{u}_{n-1}, \varepsilon_n\bar{u})$ also satisfies (1.2) on (a, b). Now the existence of factorizations of forms (9.2), (9.3), and (9.4) follows from Theorem 1.1. The other claims in the statement follow from an inspection of the factorizations of the operator d^3/dt^3 on $(0, +\infty)$; see the next proposition. □

PROPOSITION 9.2. *All the factorizations of the operator d^3/dt^3 are of the type*

$$u''' \equiv \frac{1}{A(t)}\left[\frac{A^2(t)}{B(t)}\left(\frac{B^2(t)}{A(t)}\left(\frac{u}{B(t)}\right)'\right)'\right]' \tag{9.5}$$

where $B(t)$ denotes, apart from a nonzero constant factor, a generic solution to $u'''=0$ strictly positive on the chosen open interval $\mathcal{T}$

$$B(t)=c_0+c_1t+c_2t^2 \quad (c_i=\text{constant}); \tag{9.6}$$

while $A(t)$ is defined by

$$A(t)=(\beta c_0-\alpha c_1)+2(\gamma c_0-\alpha c_2)t+(\gamma c_1-\beta c_2)t^2 \tag{9.7}$$

and depends on three arbitrary constants α, β, γ which, when $B(t)$ has been fixed, are subject to the restriction that $A(t)>0$ on $\mathcal{T}$.

In particular we explicitly mention the following factorizations on $(0,+\infty)$:

(1) *Apart from positive constant factors all the factorizations on $(0,+\infty)$ with $B(t)\equiv 1$ are*

$$u'''\equiv((u')')', \tag{9.8}$$

$$u'''\equiv(t+c)^{-1}[(t+c)^2((t+c)^{-1}u')']', \quad t>0 \quad (c=\text{constant}\geqq 0). \tag{9.9}$$

(2) *All the factorizations on $(0,+\infty)$ with $B(t)=t^2$ are (positive constant factors apart)*

$$u'''\equiv\frac{1}{t^2}\left[t^2\left(t^2\left(\frac{u}{t^2}\right)'\right)'\right]', \qquad t>0, \tag{9.10}$$

$$u'''\equiv\frac{1}{t}\left[\left(t^3\left(\frac{u}{t^2}\right)'\right)'\right]', \qquad t>0, \tag{9.11}$$

$$u'''\equiv\frac{1}{t(t+c)}\left\{(t+c)^2\left[\frac{t^3}{t+c}\left(\frac{u}{t^2}\right)'\right]'\right\}', \quad t>0 \quad (c=\text{constant}>0). \tag{9.12}$$

(3) *All the factorizations on $(0,+\infty)$ with $B(t)=t^2+c$ $(c>0)$ are (positive constant factors apart)*

$$u'''\equiv\frac{1}{t}\left\{\frac{t^2}{t^2+c}\left[\frac{(t^2+c)^2}{t}\left(\frac{u}{t^2+c}\right)'\right]'\right\}', \qquad t>0. \tag{9.13}$$

Remark. Factorizations (9.8)-(9.12) on $(0,+\infty)$ prove the claims in Proposition 9.1(2), while factorizations (9.13) prove the claim in Proposition 9.1(3).

Proof of Proposition 9.2. The proof is easy only if appropriate devices are employed. The object is to find all the possible factorizations

$$u'''\equiv p_3[p_2(p_1(p_0u)')'] \tag{9.14}$$

valid on some open interval $\mathcal{T}\subset\mathbb{R}$. If (9.14) holds on $\mathcal{T}$, then p_0 must be of the form

$$p_0(t)=\frac{1}{c_0+c_1t+c_2t^2}\equiv\frac{1}{B(t)} \tag{9.15}$$

where $B(t)>0$ on $\mathcal{T}$. Once p_0 has been chosen, p_1 must be such that the function $[1/p_0(t)]\int_{t_0}^{t}d\tau/p_1(\tau)$, with t_0 arbitrarily fixed on $\mathcal{T}$, is a solution to $u'''=0$. Hence

$$\frac{1}{p_0(t)}\int_{t_0}^{t}\frac{d\tau}{p_1(\tau)}=\alpha+\beta t+\gamma t^2$$

and

$$\begin{aligned}\frac{1}{p_1(t)}&=[(B(t))^{-1}(\alpha+\beta t+\gamma t^2)]'\\&=-(B(t))^{-2}B'(t)(\alpha+\beta t+\gamma t^2)+(B(t))^{-1}(\beta+2\gamma t)=(B(t))^{-2}A(t)\end{aligned}$$

whence

$$p_1(t)=\frac{B^2(t)}{A(t)}. \tag{9.16}$$

Once p_0 and p_1 have been chosen we shall show that p_2 is uniquely determined (up to constant factors). Let us first evaluate the differential expression

$$\begin{aligned}(p_1(p_0u)')'&=\left[\frac{B^2(t)}{A(t)}\left(\frac{u}{B(t)}\right)'\right]'\\ &=(A(t))^{-2}\{[A'(t)B'(t)-A(t)B''(t)]u-A'(t)B(t)u''\\ &\qquad +A(t)B(t)u''\}.\end{aligned} \tag{9.17}$$

On the other hand, by using the explicit expressions for $B(t)$ and $A(t)$, as given by (9.6) and (9.7), we get

$$A'(t)B'(t)-A(t)B''(t)=2(\gamma c_1-\beta c_2)B(t)$$

and, substituting into (9.17), we obtain

$$(p_1(p_0u)')'\equiv(A(t))^{-2}B(t)L(u) \tag{9.18}$$

where $L(u)$ is the differential operator defined as

$$L(u)\equiv A(t)u''-A'(t)u'+2(\gamma c_1-\beta c_2)u, \tag{9.19}$$

which satisfies the relation

$$\frac{d}{dt}L(u)=A(t)u'''. \tag{9.20}$$

By using (9.18) and (9.20) we now evaluate the differential expression

$$\begin{aligned}[p_2(p_1(p_0u)')']'&=[p_2(t)(A(t))^{-2}B(t)L(u)]'\\ &=[p_2(t)(A(t))^{-2}B(t)]'L(u)+p_2(t)(A(t))^{-2}B(t)\frac{d}{dt}L(u)\\ &=[p_2(t)(A(t))^{-2}B(t)]'L(u)+p_2(t)(A(t))^{-1}B(t)u'''.\end{aligned} \tag{9.21}$$

From this we immediately infer that identity (9.14) can hold on $\mathcal{T}$ if and only if

$$p_2(t)(A(t))^{-2}B(t)=\text{constant on } \mathcal{T},$$

whence (constant factors apart)

$$p_2(t)=\frac{A^2(t)}{B(t)}. \tag{9.22}$$

To complete the proof of Proposition 9.2 a few words concerning (3) must be added. When we choose $B(t)=t^2+c$ $(c>0)$ on some interval $\mathcal{T}$ then, constant factors apart, $A(t)$ must take either the form

$$A(t)=t^2+2\lambda t-c \quad (\lambda=\text{constant}), \quad \text{or} \tag{9.23}$$

$$A(t)=t. \tag{9.24}$$

As $c>0$ the polynomial (9.23) has a positive zero for any choice of λ; hence it can never be strictly positive on $(0,+\infty)$. Thus, if we are interested in factorizations valid on $(0,+\infty)$, the only admissible choice for $A(t)$ is (9.24). □

We draw the reader's attention to the fact that the operator d^3/dt^3, as ensured by Theorem 8.3, has double C.F.s of type (II) on $(0, +\infty)$. However these C.F.s are not of type (9.5) with $B(t) = t^2 + c$ $(c > 0)$; the admissible choices for the couple $(B(t), A(t))$ are in fact a little more involved.

REFERENCES

[1] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Mathematics 220, Springer-Verlag, Berlin, 1971.

[2] G. FROBENIUS, *Ueber adjungirte lineare Differentialausdrücke*, J. für Math., 85 (1878), pp. 185-213.

[3] A. GRANATA, *Canonical factorizations of disconjugate differential operators*, SIAM J. Math. Anal., 11 (1980), pp. 160-172.

[4] ———, *On Factorizations of Selfadjoint Ordinary Differential Operators*, Proc. Amer. Math. Soc., Providence, RI, 86, 1982, pp. 260-266.

[4_{bis}] PH. HARTMAN, *Ordinary Differential Equations*, 2nd ed., Birkhäuser, Boston, 1982.

[5] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956.

[6] A. YU. LEVIN, *Non-oscillation of solutions of the equation* $x^{(n)} + p_1(t)x^{(n-1)} + \cdots + p_n(t)x = 0$, Uspekhi Mat. Nauk, 24 (1969), pp. 43-96. (In Russian.) Russian Math Surveys, 24 (1969), pp. 43-99. (In English.)

[7] G. PÓLYA, *On the mean-value theorem corresponding to a given linear homogeneous differential equation*, Trans. Amer. Math. Soc., 24 (1922), pp. 312-324.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

014

ON THE CONVERGENCE OF INTERPOLATED ITERATION METHODS*

ERICH NOVAK†

Abstract. Convergence proofs are presented for the two iterations $f_{n+1}(x)=\alpha Af_n(x)+(1-\alpha)f_n(x)$, $0<\alpha\leqq 1$ and $f_{n+1}(x)=(Af_n(x))^{\beta}\cdot f_n(x)^{1-\beta}$, $0<\beta\leqq 1$. Here A are special integral operators arising in the nonlinear elastic deformation of circular membranes and plates.

Three problems are studied. In the first the fixed points $Af=f$ are identical with the solutions of the circular membrane problem for uniform lateral load and (dimensionless) radial edge tension S, where it is shown that both methods converge, with arbitrary choice of the starting function $f_0\geqq S$, provided α and β are suitably chosen.

The second problem concerns annular elastic membranes under the action of axisymmetric surface load and several types of edge boundary conditions. It is shown how the first iteration method can be used to obtain statements on the existence of positive solutions.

In the third problem, $Af=f$ represents the solutions of the circular plate bending problem for uniform lateral load. A conjecture concerning the first iteration method due to Keller and Reiss is proved.

Key words. approximation of fixed points, nonlinear integral equations, nonlinear elasticity

AMS(MOS) subject classifications. 45G10, 47H17, 73C50

1. Introduction. Consider the iteration method $x_{n+1}=f(x_n)$ for finding a fixed point of f. The condition $|f'(x)|\leqq R<1$ is sufficient for convergence of the iteration method. Ostrowski [8] showed that when $1>R\geqq f'(x)\geqq r$, then the interpolated iteration

$$x_{n+1}=\alpha f(x_n)+(1-\alpha)x_n \tag{I}$$

converges for the (optimal) choice $\alpha=2/(2-r-R)$ to the unique fixed point of f.

In their work on nonlinear bending of a circular elastic plate, Keller and Reiss [6] proposed the same iteration for solving a nonlinear integral equation $Af=f$ equivalent to the plate bending boundary value problem, that is,

$$f_{n+1}=\alpha Af_n+(1-\alpha)f_n. \tag{I'}$$

They argued heuristically that (I') can be expected to converge if the spectral radius of $\alpha A'f+(1-\alpha)I$ is less than 1, where I is the identity and $A'f$ is the Fréchet derivative of the operator A at the point f. Their conjecture that (I') always converges if α is suitably chosen was demonstrated numerically by extensive computations [6], but a rigorous proof has never been obtained. We remark, however, that the local convergence of (I') easily follows from the results of Kitchen [7]. The iteration (I') has also been applied successfully in [9] and [10] to the simpler problems of nonlinear deformation of circular and annular elastic membranes.

In a different context, Amann [1] considered the iteration (I'), where A is an operator that corresponds to certain Hammerstein type equations. Prior to Amann, the iteration (I') has been used by Zarantonello [11].

THEOREM 1 (Amann [1]). *Let H be a real Hilbert space, $F: H\to H$ Lipschitz continuous in each bounded set and monotone, i.e., $(Ff-Fg, f-g)\geqq 0$ for $f, g\in H$ and let $K: H\to H$ be linear, continuous, self-adjoint, and positive, i.e., $(Kf, f)\geqq 0$ for $f\in H$. Then the equation*

$$f+KFf=g \quad \text{or} \quad f=g-KFf:=Af, \tag{1}$$

* Received by the editors April 21, 1986; accepted for publication (in revised form) August 6, 1987.

† Universität Erlangen-Nürnberg, Mathematisches Institut, Bismarckstrasse 11/2, D-8520 Erlangen, Federal Republic of Germany.

with $g \in H$, has a unique solution $f^* \in H$. Furthermore, the iteration method

$$f_{n+1} = \alpha A f_n + (1-\alpha) f_n, \tag{I'}$$

$f_0 \in H$, converges to f^* for all $\alpha > 0$ with $0 < \alpha < \alpha_0$ provided α_0 is chosen sufficiently small.

2. The circular membrane under uniform load.

2.1. The convergence of the iteration (I'). Consider a circular elastic membrane subject to a uniform normal pressure and radial edge traction within the nonlinear Föppl-Hencky theory of small finite deflection. Then the radial stress is determined by the following boundary value problem:

$$y'' + 3y'/x + 2/y^2 = 0 \qquad (0 < x < 1) \tag{2}$$

and

$$y'(0) = 0, \qquad y(1) = S > 0.$$

These equations are equivalent to the integral equation

$$f(x) = S + \int_0^1 k(x, t)\varphi(t, f(t))\, dt \tag{3}$$

where $\varphi(t, u) = u^{-2}$ and

$$k(x, t) = \begin{cases} (x^{-2} - 1)t^3 & \text{for } t \leqq x, \\ (t^{-2} - 1)t^3 & \text{for } t > x. \end{cases}$$

For more details concerning these equations, see Dickey [3], Callegari and Reiss [2], and Weinitschke [10]. Clearly each solution f_S^* of (2) or (3) satisfies $f_S^* \geqq S$. Hence (3) is equivalent to

$$f + KFf = g \quad \text{or} \quad f = g - KFf := Af \tag{1}$$

where $g = S$, $Ff(t) = -\text{Max}\,(S, f(t))^{-2}$, and $Kh(x) = \int_0^1 k(x, t)h(t)\, dt$. Applying Theorem 1 to the Hilbert space H which is given by the weighted L_2-norm $\|h\| = (\int_0^1 h(x)^2 x^3\, dx)^{1/2}$, we obtain the following result.

THEOREM 2. *The boundary value problem* (2) *has a unique solution f_S^* and the iterative method*

$$f_{n+1} = \alpha A f_n + (1-\alpha) f_n, \tag{I'}$$

$f_0 \in H$, converges for all $\alpha > 0$ which are sufficiently small.

Next we propose to replace the iterative method (I') *by*

$$f_{n+1} = (Af_n)^\beta \cdot f_n^{1-\beta}, \qquad f_0 \geqq S, \tag{II}$$

where $\beta > 0$ is chosen sufficiently small.

We shall prove that this method also converges to the solution f_S^*. Numerical calculations of Weinitschke [10] show that (II) converges faster then (I'), especially for small values of $S > 0$.

2.2. Some preliminary lemmas.

LEMMA 1. *Let $S > 0$ and $M_S = \{f \in C[0, 1] \mid S \leqq f \leqq S + 1/4S^2$, f monotone decreasing$\}$. Then $A(M_S) \subseteq M_S$.*

Proof. This is an immediate consequence of the fact that A is an antitone operator in the sense that $Af \geqq Ag$ for any $f \leqq g$.

Next consider a sequence f_n in M_S, and set $g_n = \log f_n$. Then the following statements are equivalent:

(i) f_n is uniformly convergent, $\lim f_n = f$;

(ii) g_n is uniformly convergent, $\lim g_n = g = \log f$.

This implies that the iteration (II) is equivalent to $\log f_{n+1} = \beta \log Af_n + (1-\beta)\log f_n$, which can be written as

$$g_{n+1} = \beta \tilde{A} g_n + (1-\beta) g_n, \tag{II'}$$

where $\tilde{A}g = \log(A(\exp g))$. We find again that $\tilde{A}(\tilde{M}_S) \subseteq \tilde{M}_S$, if we define

$$\tilde{M}_S = \left\{ f \in C[0,1] \,|\, \log S \leqq f \leqq \log\left(S + \frac{1}{4S^2}\right), f \text{ monotone decreasing} \right\}.$$

Calculating the Fréchet derivative of $\tilde{A}$, we find

$$(\tilde{A}'g)h(x) = -2\frac{\int_0^1 k(x,t)\varphi(t, \exp g(t))h(t)\,dt}{S + \int_0^1 k(x,t)\varphi(t, \exp g(t))\,dt}.$$

LEMMA 2. *Let $g \in \tilde{M}_S$. Then we have the estimate*

$$\|\tilde{A}'g\| \leqq 2 \cdot (1+4S^3)^{-1},$$

where $\|\cdot\|$ is the operator norm with respect to the sup-norm in $C[0,1]$.

Proof. Calculating $\|\tilde{A}'g\|$, we find

$$\|\tilde{A}'g\| = \sup_{x\in[0,1]} \frac{2\int k(x,t)\varphi(t, \exp g(t))\,dt}{S + \int k(x,t)\varphi(t, \exp g(t))\,dt}$$

$$= \frac{2 \cdot \int (1-t^2) t (\exp g(t))^{-2}\,dt}{S + \int (1-t^2) t (\exp g(t))^{-2}\,dt}.$$

Since $\|\tilde{A}'g\|$ is maximal for $g = \log S$, we have $\|\tilde{A}'g\| \leqq \|\tilde{A}'(\log S)\| = 2 \cdot (1+4S^3)^{-1}$.

2.3. The convergence of $f_{n+1} = Af_n$. An immediate consequence of Lemma 2 is the following theorem.

THEOREM 3. *The iteration $f_{n+1} = Af_n$ converges for all S with $4S^3 > 1$, that is, $S > 0.63$.*

This result is a slight improvement of [3] mentioned earlier. However, it is known from [10] that the simple iteration $f_{n+1} = Af_n$ converges even for $S > 0.5609$. This improvement was achieved by replacing the sup-norm by a weighted sup-norm. Theorem 3 says that A is a contraction with respect to the metric

$$d(f,h) = \sup_{x\in[0,1]} |\log f(x) - \log h(x)|,$$

if $4S^3 > 1$. We may also improve the result of Theorem 3 by using a weighted norm for $\tilde{A}$

$$\|h\| := \sup_{x\in[0,1]} \left|\frac{h(x)}{w(x)}\right|, \qquad w(x) > 0,$$

which amounts to taking $\sup |(\log f(x) - \log h(x))/w(x)|$ as a metric for A. Computing the operator norm of $\tilde{A}'g$ with respect to the weighted norm we obtain with $\exp g = f \in M_S$, $g \in \tilde{M}_S$,

$$\|\tilde{A}'g\|_w = \sup_{x\in[0,1]} \int \frac{2k(x,t)\varphi(t,f(t))w(t)}{w(x)(S + \int k(x,t)\varphi(t,f(t))\,dt)}\,dt.$$

Since the function $af(t)^{-2}/(b+cf(t)^{-2})$, $a, b, c>0$ is increasing with decreasing f, we have

$$\|\tilde{A}'g\|_w \leqq \|\tilde{A}' \log S\|_w = \sup_{x\in[0,1]} \frac{8}{w(x)(4S^3+1-x^2)} \int_0^1 k(x,t)w(t)\,dt.$$

Similarly as in [10], we attempt to find an optimal weight w so as to make $\|\tilde{A}'(\log S)\|<1$ for $S>S_0>0$. Thus we wish to have

$$\int_0^1 k(x,t)w(t)\,dt = w(x)\left(\delta - \frac{1}{8}x^2\right), \qquad \delta := \frac{1}{8}(4S_0^3+1).$$

We omit the calculations which lead to the value $\delta = 0.1842$ and $\varepsilon_0 = 0.491$. Thus we arrive to the following result.

THEOREM 3A. *The simple iteration $f_{n+1} = Af_n$ converges for all $S > 0.491$.*

2.4. The convergence of the iteration (II). We consider operators of the form

$$Lf(x) = \int G(x,t)f(t)\,dt$$

where

$$G(x,t) = \begin{cases} b(x)c(t) & \text{for } x \leqq t, \\ b(t)c(x) & \text{for } x > t, \end{cases}$$

with the following assumptions: $b, c \in C^1[0,1]$ are nonnegative and $c(x) = g(x)b(x)$, with a nonincreasing function g. We need the following lemma.

LEMMA 3. *The operator L is positive-semidefinite, that is, $(Lh, h) \geqq 0$ for all $h \in L_2[0,1]$.*

Proof. Let h be a differentiable function and set $B(x) = \int_0^x b(t)h(t)\,dt$. It follows that

$$\int_0^1 \int_0^1 G(x,t)h(x)h(t)\,dt\,dx = 2\int_0^1 \int_0^x c(x)b(t)h(x)h(t)\,dt\,dx$$
$$= 2\int_0^1 g(x)b(x)h(x)B(x)\,dx,$$

and since $(\frac{1}{2}B^2)' = Bbh$, we obtain

$$(Lh, h) = g(1)B^2(1) - \int_0^1 g'(x)B^2(x)\,dx \geqq 0.$$

Now we consider a weighted L_2-norm

$$\|h\| = \left(\int_0^1 h(x)^2 p(x)\,dx\right)^{1/2},$$

where p is a positive weight. If the mapping $L = \tilde{A}'g$, $g \in \tilde{M}_S$ is given by $Lh(x) = \int G(x,t)h(t)\,dt$, then $L^*h(x) = \int G(t,x)p(t)\colon p(x)h(t)\,dt$ and $\|L\| = (\lambda_{\max}(LL^*))^{1/2} \leqq (\iint G(x,t)^2 p(x)\colon p(t)\,dt\,dx)^{1/2}$. The quadratic form that is related to L is given by

$$Lh \cdot h = \int_0^1 \int_0^1 G(x,t)h(t)h(x)p(x)\,dt\,dx.$$

If $\beta \in]0, 1]$ and $L_\beta = \beta L + (1-\beta)I$, then

$$\|L_\beta\| \leqq (\beta^2\|L\| + (1-\beta)^2 + 2\beta(1-\beta) \cdot C)^{1/2},$$

where

$$C = \sup_{h \neq 0} Lh \cdot h / h \cdot h.$$

If $C < 1$ then L_β is a contraction for all $\beta > 0$ that are sufficiently small. With $L = \tilde{A}'g$, $f(t) = (\exp g(t))^{-2}$ and $l(x) = (S + \int k(x, t)(\exp g(t))^{-2}\, dt)^{-1}$ we have

$$G(x, t) = \begin{cases} -2l(x)(1-t^2)tf(t), & x \leqq t, \\ -2(1-x^2) : x^2 l(x) f(t) t^3, & x > t. \end{cases}$$

The mapping $L = \tilde{A}'g$ has the form

$$\begin{aligned} Lh(x) &= -2\int_0^x b_2(t)c_2(x)h(t)\, dt - 2\int_x^1 b_1(x)c_1(t)h(t)\, dt \\ &= L_2h(x) + L_1h(x), \end{aligned}$$

where

$$\begin{aligned} b_2(t) &= f(t)t^3, & c_2(x) &= (1-x^2) : x^2 l(x), \\ b_1(x) &= l(x), & c_1(t) &= (1-t^2)tf(t). \end{aligned}$$

The related quadratic functional may be written as $Lh \cdot h = L_2h \cdot h + L_1h \cdot h$. Using Lemma 3 we see that $L_2h \cdot h$ is negative-semidefinite, if $b_2/(c_2p)$ is monotone increasing.

LEMMA 4. *In the case $p(x) = x^5$ the quadratic functional $Q(h) = L_2h \cdot h$ is negative-semidefinite for all $S > 0$ and $g \in M_S$.*

Proof. We have to show that $H(x) = b_2(x)/(c_2(x)p(x))$ is increasing

$$H(x) = \frac{f(x)x^5(S + \int k(x, t)f(t)\, dt)}{(1-x^2)x^5} = \frac{f(x)(S + \int k(x, t)f(t)\, dt)}{1-x^2}$$

yields

$$H'(x) = -2x^{-3}\int_0^x t^3 f(t)\, dt + \frac{2x}{(1-x^2)^2}\int_x^1 (1-t^2)tf(t)\, dt,$$

which is nonnegative if $f \geqq 0$ is increasing. As the function

$$N(x) = b_1(x)p(x)/c_1(x) = \frac{x^4}{f(x)(1-x^2)(S + \int k(x, t)f(t)\, dt)}$$

is not increasing we cannot conclude that $L_1h \cdot h$ is negative-semidefinite. We assume that $f(x) = (S + \int k(x, t)r(t)\, dt)^{-2}$ with a positive increasing function r. As in Lemma 4, it follows that $f(x) \cdot (1-x^2)^2$ is decreasing. Since $(S + \int k(x, t)f(t)\, dt)$ is decreasing too, the function N is increasing at least in those subintervals where $x^4(1-x^2)$ is increasing. Hence N is increasing in $[0, \sqrt{2/3}]$. Now we approximate L_1 by a mapping $\tilde{L}_1$ such that $\tilde{Q}(h) = \tilde{L}_1h \cdot h$ is negative-semidefinite. We define $\tilde{L}_1$ by

$$\tilde{L}_1h(x) = -2\int_x^1 \tilde{b}_1(x)c_1(t)h(t)\, dt,$$

where

$$\tilde{b}_1(x) = b_1(x)$$

for $x \in [0, \sqrt{2/3}]$ and

$$\tilde{b}_1(x) = c_1(x)\colon p(x) \cdot \left(\frac{b_1 p}{c_1}\left(\sqrt{\frac{2}{3}}\right)\right) \quad \text{for } x > \sqrt{\frac{2}{3}}.$$

Clearly $\tilde{b}_1 p / c_1$ is increasing on $[0, 1]$ and $\tilde{Q}$ is negative-semidefinite by Lemma 3. Furthermore, we have

$$c = \sup_{h \neq 0} \frac{Lh \cdot h}{h \cdot h} \leqq \| L_1 - \tilde{L}_1 \|.$$

Hence

$$c^2 \leqq 4 \cdot \int_{\sqrt{2/3}}^1 \int_x^1 (\tilde{b}_1(x) - b_1(x))^2 c_1(t)^2 p(x)\colon p(t)\, dt\, dx.$$

The integral is maximal for $f = \text{const}$ and $\varepsilon \to 0$, and we get

$$c^2 \leqq 4 \cdot \int_{\sqrt{2/3}}^1 \int_x^1 \left(\frac{4}{1-x^2} - \frac{16(1-x^2)}{x^4}\right)^2 (1-t^2)^2 t^2 x^5\colon t^5\, dt\, dx.$$

A computation of the integral yields about 0.6, and thus we have proved the following theorem.

THEOREM 4. *Assume that $\beta > 0$ is sufficiently small. Then the iteration method*

$$f_{n+1}(x) = (Af_n(x))^{\beta} \cdot f_n(x)^{(1-\beta)},$$

$S > 0$, $f_0 \geqq S$, converges to the unique fixed point f_S^* of A.

The convergence follows from the contraction lemma with the metric

$$d(f, g) = \left(\int_0^1 (\log f(x) - \log g(x))^2 x^5\, dx\right)^{1/2}.$$

3. Nonlinear boundary value problems for the annular membrane. We consider here an annular elastic membrane under the action of axisymmetric surface loads and uniform radial edge stresses or displacements, again within the Föppl-Hencky theory. This leads us to consider four different boundary value problems for the nonlinear differential equation

$$y'' + 3y'/x + 2R^2(x)/y^2 = 0, \qquad 0 < a < x < 1,$$

where R is nondecreasing with $R(a) = 0$. We refer to [4], [5], or [9] for a more complete description of this problem. By means of appropriate Green functions, these problems can be written as integral equations of the following form:

$$f(x) = g(x) - \int_0^1 k(x, t)\varphi(t, f(t))\, dt. \tag{4}$$

We are interested in positive solutions of (4) and observe that in all four cases the following conditions are fulfilled:

(i) $\varphi\colon [0, 1] \times [\varepsilon, \infty[\to \mathbb{R}$ is Lipschitz continuous for every $\varepsilon > 0$;

(ii) $\varphi(t, \cdot)\colon \mathbb{R}^+ \to \mathbb{R}$ is nondecreasing for each $t \in [0, 1]$;

(iii) The kernel k is continuous, symmetric, and positive semidefinite, i.e.,

$$\int_0^1 \int_0^1 k(x, t)h(x)h(t)\, dt\, dx \geqq 0 \quad \text{for all } h \in L_2[0, 1];$$

(iv) g is continuous.

Under these conditions, a positive solution of (4) does not always exist. Although in many cases it can be decided by means of a priori estimates whether a positive solution exists, such a decision is not always possible (see [5]). The following theorem shows that the positive solution of (4) (if it exists) is unique. For a different proof, see [4]. Moreover we give a method how to decide whether a positive solution exists and how to construct it. We always assume that the conditions (i)-(iv) are satisfied.

THEOREM 5. *Let* $\varepsilon > 0$ *and* $A_\varepsilon : L_2[0,1] \to L_2[0,1]$ *be defined by*

$$A_\varepsilon f(x) = g(x) - \int_0^1 k(x,t)\varphi(t, \operatorname{Max}(f(t), \varepsilon))\, dt.$$

Then

(i) A_ε *has a unique fixed point* f_ε (*which is continuous*);

(ii) f_ε *can be obtained by means of the iteration*

$$f_{n+1} = \alpha A_\varepsilon f_n + (1-\alpha) f_n, \tag{I'}$$

where f_0 *is arbitrary and* $\alpha > 0$ *is sufficiently small*;

(iii) *Equation* (4) *has a positive solution* f^* *with* $\inf f^* \geqq c$ *if and only if the fixed point* f_c *of* A_c *fulfills* $f_c \geqq c$. *In this case we have* $f^* = f_\varepsilon$ *for all* $\varepsilon \leqq c$;

(iv) *The positive solution* f^* *of* (4) *is unique* (*if it exists*).

Proof. Statements (i) and (ii) follow from Theorem 1 with $H = L_2[0,1]$ and $A = A_\varepsilon$. The continuity of f_ε follows from the fact that g and k are continuous. If f^* is a solution of (4) with $f^* \geqq c > 0$, then $A_\varepsilon f^* = f^*$ for all $\varepsilon \leqq c$. Hence we have $f^* = f_\varepsilon$ for $\varepsilon \leqq c$ and therefore $f_c \geqq c$. If, on the other hand, $c > 0$ and $f_c \geqq c$, then we have $A_\varepsilon f_c = f_c$ for all $\varepsilon \leqq c$ and $f_c = f_\varepsilon$ ($\varepsilon \leqq c$) is a positive solution of (4). Assume that f_1 and f_2 are positive solutions of (4). Then f_1 and f_2 are solutions of $A_\varepsilon f = f$ for $\varepsilon \leqq \operatorname{Min}(\inf f_1, \inf f_2)$, and hence $f_1 = f_2$.

4. Iterative solution for the nonlinear bending of circular plates. We study a thin circular elastic plate subjected to uniform lateral pressure within the nonlinear theory of von Kármán. According to Keller and Reiss [6], the resulting boundary value problems can be reduced to an equivalent integral equation of the form

$$f(x) = -\tfrac{1}{2} G_1(f \cdot G_2(f^2))(x) + g(x) := Af(x), \tag{5}$$

where

$$G_i f(x) = -\int_0^1 g_i(x,t) f(t)\, dt, \qquad i = 1, 2$$

with

$$g_i(x,t) = \begin{cases} \dfrac{-1}{2\mu_i}\left(\dfrac{\mu_i}{x} + x\right) t & \text{for } t \leqq x, \\ \dfrac{-1}{2\mu_i}\left(\dfrac{\mu_i}{t} + t\right) x & \text{for } t > x \end{cases}$$

and

$$g(x) = \gamma x(1 - x^2).$$

The constants μ_i and γ satisfy $\mu_1 = -1$, $\mu_2 > 0$, and $\gamma > 0$, where γ is proportional to the applied pressure.

The existence of a unique solution of (5) was proved in [6] for a restricted range of γ. More precisely, it was shown that the simple iteration $f_{n+1} = Af_n$ converges for $\gamma \leqq \gamma_0$ but diverges for $\gamma \geqq \gamma_1$, where numerical estimates for γ_0 and γ_1 were given. The following theorem proves the conjecture of Keller and Reiss [6], described in the Introduction, and at the same time yields the existence of a unique solution for *arbitrary* large positive values of γ.

THEOREM 6. *Equation* (5) *has a unique solution* f^* *and the iteration method*

$$f_{n+1} = \alpha A f_n + (1-\alpha) f_n, \tag{I'}$$

$f_0 \in H = L_2[0, 1]$, $\alpha > 0$ *sufficiently small, converges and* $\lim_{n\to\infty} f_n = f^*$.

Proof. We define $F: H \to H$ by $Ff(x) = f(x) \cdot G_2(f^2)(x)$. Let $f, h \in H = L_2[0, 1]$. We have

$$\begin{aligned} \|G_2(f^2) - G_2(h^2)\|_\infty &= \|G_2(f^2 - h^2)\|_\infty \leqq c \cdot \|f^2 - h^2\|_1 \\ &= c \cdot (|f-h|, |f+h|) \leqq c \cdot \|f-h\|_2 \cdot \|f+h\|_2 \end{aligned}$$

for some $c > 0$, and hence

$$\begin{aligned} \|Ff - Fh\|_2 &= \|fG_2(f^2) - hG_2(h^2)\|_2 \\ &\leqq \|fG_2(f^2) - fG_2(h^2)\|_2 + \|fG_2(h^2) - hG_2(h^2)\|_2 \\ &\leqq \|f\|_2 \cdot c\|f-h\|_2 \cdot \|f+h\|_2 + \|f-h\|_2 \cdot c \cdot \|h\|_2^2 \\ &= c \cdot \|f-h\|_2 \cdot (\|f\|_2 \cdot \|f+h\|_2 + \|h\|_2^2). \end{aligned}$$

Thus F is Lipschitz continuous on bounded sets of H. Because of

$$F'f(h)(x) = h(x)G_2(f^2)(x) + f(x)G_2(2fh)(x),$$

we get

$$(F'f(h), h) = \int_0^1 h^2(x)G_2(f^2)(x)\,dx + \int_0^1 f(x)h(x)G_2(2fh)(x)\,dx.$$

It follows from Lemma 3 that G_2 is positive, and hence $(F'f(h), h) \geqq 0$ and F is a monotone operator. Since G_1 is (again by Lemma 3) positive we can apply Theorem 1 and the proof is complete.

Acknowledgment. I would like to express my thanks to Professor H. J. Weinitschke for valuable comments concerning this paper. He initiated this work during my stay at the Institut für Angewandte Mathematik, Erlangen, and also helped with the final version of this paper.

REFERENCES

[1] H. AMANN, *Über die Existenz und iterative Berechnung einer Lösung der Hammerstein'schen Glcichung*, Aequationes Math., 1 (1968), pp. 242-266.

[2] A. J. CALLEGARI AND E. L. REISS, *Non-linear boundary value problems for the circular membrane*, Arch. Rational Mech. Anal., 31 (1968), pp. 390-400.

[3] R. W. DICKEY, *The plane circular elastic surface under normal pressure*, Arch. Rational Mech. Anal., 26 (1967), pp. 219-236.

[4] H. GRABMÜLLER AND E. NOVAK, *Nonlinear boundary value problems for the annular membrane: a note on uniqueness of positive solutions*, J. Elasticity, 17 (1987), pp. 279-284.

[5] ———, *Nonlinear boundary value problems for the annular membrane: new results on existence of positive solutions*, Math. Method Appl. Sci., to appear.

[6] H. B. KELLER AND E. L. REISS, *Iterative solutions for the non-linear bending of circular plates*, Comm. Pure Appl. Math., 11 (1958), pp. 273-292.

[7] J. W. KITCHEN, JR., *Concerning the convergence of iterates to fixed points*, Studia Math., 27 (1966), pp. 247-249.

[8] A. M. OSTROWSKI, *Solution of Equations in Euclidean and Banach Spaces*, Academic Press, New York, London, 1973.

[9] H. J. WEINITSCHKE, *On axisymmetric deformations of nonlinear elastic membranes*, in Mechanics Today 5, S. Nemat-Nasser, ed., Pergamon Press, Oxford, New York, 1980, pp. 523-542.

[10] ———, *On finite displacements of circular elastic membranes*, Math. Methods Appl. Sci., 9 (1987), pp. 76-98.

[11] E. H. ZARANTONELLO, *Solving functional equations by contractive averaging*, Technical Report 160, Mathematics Research Center, University of Wisconsin, Madison, WI, 1960.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

015

RICCATI MATRIX DIFFERENCE EQUATIONS AND DISCONJUGACY OF DISCRETE LINEAR SYSTEMS*

CALVIN D. AHLBRANDT† AND JOHN W. HOOKER‡

Abstract. Disconjugacy criteria analogous to well-known results of W. T. Reid for linear differential systems are obtained here for the linear vector difference equation

$$-\Delta(C_{n-1}\Delta x_{n-1})+A_n x_n=0,$$

where A_n and C_n are real symmetric (or complex hermitian) matrices. The coefficients C_n are assumed to be nonsingular, but C_n and A_n are not assumed to be positive definite. Disconjugacy is defined in terms of the concept of conjugate intervals. A discrete Riccati matrix operator is defined, which plays a central role in the discussion of disconjugacy. This work extends and generalizes earlier work of the authors.

Key words. difference equation, linear system, disconjugacy, discrete Riccati matrix equation

AMS(MOS) subject classifications. 39A10, 39A12, 34C10

1. Introduction.

Consider a linear vector difference equation

$$l[x]_n=-\Delta(C_{n-1}\Delta x_{n-1})+A_n x_n=0, \qquad n=1,2,3,\cdots, \tag{1.1}$$

where $\{A_n\}$ and $\{C_n\}$ are given sequences of $r\times r$ real symmetric (or complex hermitian) matrices, x_n is an $r\times 1$ vector, and Δ is the forward difference operator $\Delta x_n=x_{n+1}-x_n$. Discrete-time linear systems and related discrete matrix Riccati equations arise in discrete linear optimal control and filtering problems (cf. Vaughan [24], Kalman [14], and Kwakernaak and Sivan [15]). This paper extends results obtained by the authors in [2] and [3] for the scalar case of (1.1) with real coefficients a_n and c_n, $c_n>0$. Some of the arguments used in [2] for the scalar case, which depend on the assumption $c_n>0$, do not generalize to the vector equation (1.1), so different techniques are used here, and we do *not* assume that C_n is positive definite. Principal solutions of (1.1) were discussed by the authors in [1] under the assumption that A_n and C_n are positive definite for all n.

We will make use of the related matrix difference equation

$$L[X]_n=-\Delta(C_{n-1}\Delta X_{n-1})+A_n X_n=0, \tag{1.2}$$

where X_n is an $r\times r$ matrix for each n.

Recent oscillation results concerning the vector or scalar case of (1.1) appear in [2], [3], [7], [9]–[13], [16], [17], [23], where (1.1) is sometimes given in the alternative three-term recurrence relation form

$$l[x]_n=-C_n x_{n+1}-C_{n-1}x_{n-1}+B_n x_n=0 \tag{1.3}$$

with $B_n=C_n+C_{n-1}+A_n$, $n=1, 2, 3,\cdots$. A general discussion of the scalar case of (1.3), including basic results on oscillation and boundary value problems, has been carried out by Atkinson [4] and Fort [6].

Similarly, (1.2) may be expressed as

$$L[X]_n=-C_n X_{n+1}-C_{n-1}X_{n-1}+B_n X_n=0. \tag{1.4}$$

* Received by the editors August 4, 1986; accepted for publication (in revised form) October 10, 1987.
† Department of Mathematics, University of Missouri, Columbia, Missouri 65211.
‡ Department of Mathematics, Southern Illinois University, Carbondale, Illinois 62901.

For integers M and N, with $0<M<N$, a *solution* of (1.1) on the interval $[M, N]$ is a sequence $x=\{x_n\}$ of real or complex $r\times 1$ vectors, defined at least for $n=M-1, \cdots, N+1$, and satisfying (1.1) for $n=M, \cdots, N$.

In § 2 below, the concept of disconjugacy is defined for (1.1), along with the notions of self-conjoined solutions of (1.1) and prepared or isotropic solutions of (1.2).

In § 3 we then proceed to formulate and prove a discrete version of the well-known variational results of Reid [19], [20], which give several conditions equivalent to disconjugacy. Extensive use will be made of solutions of the related discrete Riccati equation

$$R[W]_n=0, \tag{1.5}$$

where $R[W]$ is the discrete matrix Riccati operator defined by

$$R[W]_n=-W_{n+1}+A_n+W_n(W_n+C_{n-1})^{-1}C_{n-1}.$$

This operator and the operator $L[X]$ defined by (1.2) will be seen to be related by the matrix identity (where $'$ denotes conjugate transpose)

$$X'_nL[X]_n=X'_nR[W]_nX_n$$

analogous to that which holds for differential systems (cf. Reid [19, p. 740], [20, p. 667]). For an extensive discussion of the relationship between Riccati matrix operators and associated linear systems in the continuous case, see Reid [22].

2. Definitions and preliminary results. We include here for ready reference the following elementary properties of the forward difference operator:

$$\Delta(a_nb_n)=a_n\Delta b_n+(\Delta a_n)b_{n+1}=a_{n+1}\Delta b_n+(\Delta a_n)b_n, \tag{2.1}$$

$$\sum_M^N \Delta a_n=a_{N+1}-a_M, \tag{2.2}$$

and the resulting "summation-by-parts" formula

$$\sum_M^N a_n\Delta b_n=a_Nb_{N+1}-a_{M-1}b_M-\sum_M^N(\Delta a_{n-1})b_n. \tag{2.3}$$

Given a sequence $C=\{C_n\}$ of hermitian matrices, we introduce a "bracket function" defined for complex vector or matrix sequences u and v as

$$\begin{aligned}\{u, v\}_n&=u'_nC_{n-1}\Delta v_{n-1}-(C_{n-1}\Delta u_{n-1})'v_n\\&=u'_{n-1}C_{n-1}v_n-u'_nC_{n-1}v_{n-1}.\end{aligned} \tag{2.4}$$

Note that the bracket function satisfies

$$\{u, v\}=-\{v, u\}' \tag{2.5}$$

as well as the following property.

LEMMA 2.1. *Let A and C be hermitian sequences. Then for any vector solutions u and v of* (1.1), $\{u, v\}_n$ *is constant, and for any matrix solutions U and V of* (1.2), $\{U, V\}_n$ *is constant.*

Proof. By use of the assumptions that A_n and C_n are hermitian, it is readily verified for solutions of (1.1) and (1.2) that $\Delta\{u, v\}_n\equiv 0$ and $\Delta\{U, V\}_n\equiv 0$, respectively.

DEFINITION. Vector sequences u and v are *conjoined* (Reid [19]) if $\{u, v\}_n\equiv 0$. A vector sequence u is *self-conjoined* if $\{u, u\}_n\equiv 0$, i.e., if

$$u'_{n-1}C_{n-1}u_n=u'_nC_{n-1}u_{n-1}, \qquad n=1, 2, \cdots. \tag{2.6}$$

Similarly, a matrix sequence X is called *prepared* (Hartman [8]), or *isotropic* (Coppel [5]) if $\{X, X\}_n \equiv 0$, or, equivalently, if $X'_{n-1}C_{n-1}X_n$ is hermitian for every n.

Note 2.1. If the coefficients A_n and C_n in (1.1) are real symmetric matrices, then $\{u, u\}_n$ is real for every real solution of (1.1). Thus every real vector solution is self-conjoined by (2.5) and, in particular, satisfies (2.6). The identity (2.6) is used in the proof of our main theorem. This is the main reason for restricting our attention below to real solutions and real coefficients. Otherwise we must consider self-conjoined vector solutions instead of real vector solutions and symmetric matrices must be replaced by hermitian matrices.

In order to formulate our main theorem, we need a discrete analogue of the concept of disconjugacy. While this has been introduced in various ways by different authors (cf. Hartman [7] and Peterson [18]), we will use here a notion of conjugate intervals. It is natural to speak of conjugate intervals, rather than conjugate points, for a solution u of a discrete equation, since the discrete analogue of a zero of a solution of a scalar differential equation may be either a value n such that $u_n = 0$ or a pair of values $(n, n+1)$ such that $u_n u_{n+1} < 0$.

DEFINITION 2.1. For positive integers M and N, with $M < N$, (1.1) is called *disconjugate* on $[M-1, N]$ if the real interval $[M-1, N]$ contains no pair of *conjugate intervals.* Distinct real intervals $[p, p+1]$ and $[q, q+1]$ (where p and q are integers in $[M-1, N-1]$) are called *conjugate intervals* if there exists a real vector solution x of (1.1) such that

$$x'_p C_p x_{p+1} \leqq 0 \quad \text{and} \quad x'_q C_q x_{q+1} \leqq 0, \tag{2.7}$$

with $x_{p+1} \neq 0$, $x_{q+1} \neq 0$, and x_n not identically zero for n in the interior of the smallest interval containing $[p, p+1]$ and $[q, q+1]$.

3. Disconjugacy criteria. We are now in a position to state discrete analogues of several of Reid's disconjugacy criteria [19, Thm. 2.1], [20, Thm. 5.2], [20, Thm. 5.1]. In order to simplify the details in the proof below, we restrict ourselves to real vector and matrix solutions of (1.1) and (1.2), and we assume the following condition:

$$A_n \text{ and } C_n \text{ are } r \times r \text{ real symmetric matrices for all } n. \tag{3.1}$$

Our theorem could equally well be stated for complex solutions, under assumption (3.1), or for complex coefficients, i.e., A_n and C_n complex hermitian matrices. The proof is the same in outline for these cases, and we will point out some of the modifications which would be needed.

We assume also the following nonsingularity condition:

$$C_n \text{ is nonsingular for } n = 0, 1, 2, \cdots. \tag{3.2}$$

THEOREM 3.1. *Assume conditions* (3.1)-(3.2). *Let M and N be positive integers with $M < N$. Then the following conditions are equivalent*:

(i) *If u is a real vector solution of* (1.1) *on $[M, N]$ with $u'_{M-1}C_{M-1}u_M \leqq 0$ and $u_M \neq 0$, then*

$$u'_n C_n u_{n+1} > 0 \quad \text{for } n = M, \cdots, N-1.$$

(ii) *If v is a real vector solution of* (1.1) *on $[M, N]$ with $v'_{N-1}C_{N-1}v_N \leqq 0$ and $v_{N-1} \neq 0$, then*

$$v'_n C_n v_{n+1} > 0 \quad \text{for } n = M-1, \cdots, N-2.$$

(iii) *Equation* (1.1) *is disconjugate on $[M-1, N]$.*

(iv) *There exists a prepared matrix solution X of* (1.2) *on* $[M, N]$ *with*

$$X'_{n-1}C_{n-1}X_n > 0 \quad \text{for } n = M, \cdots, N.$$

(v) *There exists a sequence W of symmetric matrices, defined for* $n = M, \cdots, N+1$, *with* $W_n + C_{n-1} > 0$ *for* $n = M, \cdots, N$, *satisfying the Riccati matrix equation*

$$W_{n+1} - A_n = W_n(W_n + C_{n-1})^{-1}C_{n-1}, \qquad n = M, \cdots, N. \tag{3.3}$$

(vi) *The quadratic form J_2 defined by*

$$J_2[\eta] = \sum_M^N (\Delta\eta_{n-1})'C_{n-1}\Delta\eta_{n-1} + \eta'_n A_n \eta_n \tag{3.4}$$

is positive definite on the class of real vector sequences η with

$$\eta_{M-1} = 0 = \eta_N.$$

We note that with modifications in the proof, conditions (iv) and (v) can be replaced by the following inequality conditions:

(iv′) There exists a prepared matrix solution X satisfying the inequality $X'_n L[X]_n \geqq 0$ for $n = M, \cdots, N$, with

$$X'_{n-1}C_{n-1}X_n > 0 \quad \text{for } n = M, \cdots, N.$$

(v′) There exists a sequence W of symmetric matrices with $W_n + C_{n-1} > 0$ for $n = M, \cdots, N$, satisfying the inequality

$$W_{n+1} - A_n \leqq W_n(W_n + C_{n-1})^{-1}C_{n-1}, \qquad n = M, \cdots, N.$$

As a corollary to the theorem we also have the following result.

COROLLARY 3.1. *Each of the conditions* (i)-(vi) *above implies the strengthened Legendre condition*

$$B_n = A_n + C_n + C_{n-1} > 0 \quad \text{for } n = M, \cdots, N-1.$$

This is most easily proved by assuming condition (iv). Then equation (1.2) written in the form (1.4) yields

$$X'_n C_n X_{n+1} + X'_n C_{n-1} X_{n-1} = X'_n B_n X_n.$$

Since X satisfies condition (iv), the corollary follows immediately.

The fact that $B_n > 0$ is indeed a discrete analogue of the strengthened Legendre condition of the calculus of variations is discussed in [3, p. 14] where it is shown that the quadratic form J_2 of (3.4) may be written as

$$J_2[n] = \sum_M^N (\eta'_n B_n \eta_n - \eta'_{n-1}C_{n-1}\eta_n - \eta'_n C_{n-1}\eta_{n-1}).$$

The matrix associated with this quadratic form is the block tridiagonal matrix

$$T = \begin{bmatrix} B_M & -C_M & & \\ -C_M & & \ddots & \\ & & & -C_{N-2} \\ & & -C_{N-2} & B_{N-1} \end{bmatrix}.$$

In view of condition (vi), we are led immediately to a second corollary.

COROLLARY 3.2. *If the nonsingularity condition* (3.2) *is satisfied and T is positive definite, then all of the conditions* (i)-(vi) *hold.*

For a Sturmian comparison result along the lines of Proposition 10 of Coppel [5], we need only compare tridiagonal matrices $\tilde{T}$ to the above matrix T in order for disconjugacy to be preserved.

COROLLARY 3.3. *Suppose $\tilde{T}$ is of the same form as T with B_n and C_n replaced by $\tilde{B}_n$ and $\tilde{C}_n$, where C_n and $\tilde{C}_n$ both satisfy the nonsingularity condition* (3.2). *If the system* (1.1) *is disconjugate on* $[M-1, N]$ *and* $\tilde{T} \geqq T$, *i.e.*, $\tilde{T}-T$ *is positive semidefinite, then* $\tilde{L}[x]=0$ *is disconjugate on* $[M-1, N]$. *In particular, if* $\tilde{C}_n = C_n$ *for* $n = M, \cdots, N-2$, *and* $\tilde{B}_n \geqq B_n$ *for* $n = M, \cdots, N-1$, *then disconjugacy of* (1.1) *implies disconjugacy of* $\tilde{L}[x]=0$.

Before proceeding to the proof of the theorem, we need three lemmas.

LEMMA 3.1. *Let X be a real matrix sequence, with X_n nonsingular for $n = M-1, \cdots, N$, and let W be defined by the Riccati transformation*

$$W_n = (C_{n-1}\Delta X_{n-1})X_{n-1}^{-1}, \qquad n = M, \cdots, N+1. \tag{3.5}$$

Then $W_n + C_{n-1}$ is nonsingular and W satisfies

$$X_n' L[X]_n = X_n' R[W]_n X_n \quad \text{for } n = M, \cdots, N, \tag{3.6}$$

where $R[W]$ is the Riccati difference operator defined by

$$R[W]_n = -W_{n+1} + W_n(W_n + C_{n-1})^{-1}C_{n-1} + A_n, \qquad n = M, \cdots, N. \tag{3.7}$$

Furthermore, if X is a prepared sequence, then W_n is symmetric for $n = M, \cdots, N+1$.

Proof. Given X and W as stated, we have

$$\begin{aligned} L[X]_n &= -\Delta(C_{n-1}\Delta X_{n-1}) + A_n X_n = -C_n \Delta X_n + C_{n-1}\Delta X_{n-1} + A_n X_n \\ &= -W_{n+1}X_n + W_n X_{n-1} + A_n X_n, \qquad n = M, \cdots, N. \end{aligned} \tag{3.8}$$

Thus, for $n = M, \cdots, N$,

$$X_n' L[X]_n = X_n'(-W_{n+1} + W_n X_{n-1} X_n^{-1} + A_n)X_n. \tag{3.9}$$

Now (3.5) may be written as

$$W_n + C_{n-1} = C_{n-1} X_n X_{n-1}^{-1}, \qquad n = M, \cdots, N+1, \tag{3.10}$$

so $W_n + C_{n-1}$ is nonsingular for $n = M, \cdots, N$. From (3.10) we obtain $X_{n-1}X_n^{-1} = (W_n + C_{n-1})^{-1}C_{n-1}$. Substituting this into (3.9) yields (3.6). Equation (3.10) may also be written as

$$W_n + C_{n-1} = (X_{n-1}')^{-1}(X_{n-1}' C_{n-1} X_n)X_{n-1}^{-1}, \tag{3.11}$$

so if X is prepared then $W_n + C_{n-1}$, and hence W_n also, is symmetric for $n = M, \cdots, N+1$.

LEMMA 3.2. *If y is a vector sequence defined for $n = p-1, \cdots, q+1$, then*

$$\sum_p^q [(\Delta y_{n-1})' C_{n-1}\Delta y_{n-1} + y_n' A_n y_n] = y_q' C_q \Delta y_q - y_{p-1}' C_{p-1}\Delta y_{p-1} + \sum_p^q y_n' l[y]_n.$$

Proof. Consider the summation-by-parts formula (2.3) written in the form

$$\sum_p^q (\Delta S_{n-1})T_n = S_q T_{q+1} - S_{p-1}T_p - \sum_p^q S_n \Delta T_n.$$

Application of this formula with $S_n = y'_n$, $T_n = C_{n-1}\Delta y_{n-1}$ yields

$$\sum_{p}^{q} [(\Delta y_{n-1})' C_{n-1} \Delta y_{n-1} + y'_n A_n y_n]$$

$$= y'_q C_q \Delta y_q - y'_{p-1} C_{p-1} \Delta y_{p-1} + \sum_{p}^{q} [-y'_n \Delta(C_{n-1}\Delta y_{n-1}) + y'_n A_n y_n].$$

Since $l[y]_n = -\Delta(C_{n-1}\Delta y_{n-1}) + A_n y_n$, this proves the lemma.

LEMMA 3.3. *Suppose u is a real (or self-conjoined) solution of* (1.1) *on* $[M, N]$, *and for integers p and q satisfying $M-1 \leqq p < q < N$, let*

$$\eta_n = \begin{cases} u_n, & p+1 \leqq n \leqq q, \\ 0, & \text{otherwise.} \end{cases}$$

Then $J_2[\eta]$ as defined by (3.4) *satisfies*

$$J_2[\eta] = u'_p C_p u_{p+1} + u'_q C_q u_{q+1}. \tag{3.12}$$

(Note. If u were not real (or self-conjoined), the expression $u'_p C_p u_{p+1}$ in (3.12) would be replaced by $u'_{p+1} C_p u_p$, as in (3.17) below.)

Proof. ($p+2 \leqq q$). From the definition of η, we have

$$\begin{aligned} J_2[\eta] &= \sum_{p+1}^{q+1} [(\Delta\eta_{n-1})' C_{n-1} \Delta\eta_{n-1} + \eta' A_n \eta_n] \\ &= (\Delta\eta_p)' C_p \Delta\eta_p + \eta'_{p+1} A_{p+1} \eta_{p+1} + (\Delta\eta_q)' C_q \Delta\eta_q \\ &\quad + \sum_{p+2}^{q} [(\Delta u_{n-1})' C_{n-1} \Delta u_{n-1} + u'_n A_n u_n]. \end{aligned} \tag{3.13}$$

Also, since u is a solution of (1.1), Lemma 3.2 implies

$$\sum_{p+1}^{q} [(\Delta u_{n-1})' C_{n-1} \Delta u_{n-1} + u'_n A_n u_n] = u'_q C_q \Delta u_q - u'_p C_p \Delta u_p.$$

We rewrite this as

$$\begin{aligned} &\sum_{p+2}^{q} [(\Delta u_{n-1})' C_{n-1} \Delta u_{n-1} + u'_n A_n u_n] + (\Delta u_p)' C_p \Delta u_p + u'_{p+1} A_{p+1} u_{p+1} \\ &\quad = u'_q C_q \Delta u_q - u'_p C_p \Delta u_p. \end{aligned} \tag{3.14}$$

Then substitution from (3.14) into (3.13) yields

$$\begin{aligned} J_2[\eta] &= (\Delta\eta_p)' C_p \Delta\eta_p + \eta'_{p+1} A_{p+1} \eta_{p+1} + (\Delta n_q)' C_q \Delta\eta_q \\ &\quad + u'_q C_q \Delta u_q - u'_p C_p \Delta u_p - (\Delta u_p)' C_p \Delta u_p - u'_{p+1} A_{p+1} u_{p+1}. \end{aligned} \tag{3.15}$$

Now

$$\eta_{p+1} = u_{p+1}, \quad \Delta\eta_q = \eta_{q+1} - \eta_q = -u_q, \quad \Delta\eta_p = \eta_{p+1} - \eta_p = u_{p+1}, \tag{3.16}$$

and use of these relations in (3.15) gives us

$$J_2[\eta] = u'_q C_q u_{q+1} + u'_{p+1} C_p u_p. \tag{3.17}$$

By Note 2.1, this yields the desired result, for $q \geqq p+2$.

For the case $q=p+1$, the sum from $p+2$ to q in (3.13) vanishes, and (3.13), (3.16), and (1.3) together then imply

$$\begin{aligned} J_2[\eta] &= u'_{p+1}C_p u_{p+1} + u'_{p+1}A_{p+1}u_{p+1} + u'_{p+1}C_{p+1}u_{p+1} \\ &= u'_q B_q u_q = u'_q C_q u_{q+1} + u'_{p+1}C_p u_p, \end{aligned}$$

which completes the proof.

We now proceed to prove Theorem 3.1, in the following order: (iv)$\Rightarrow$(v)$\Rightarrow$(vi)$\Rightarrow$(i)$\Leftrightarrow$(ii)$\Rightarrow$(iii)$\Rightarrow$(iv).

Proof of Theorem 3.1. (iv)$\Rightarrow$(v). Assume that X is a prepared matrix solution of (1.2) on $[M, N]$, with

$$X'_{n-1}C_{n-1}X_n > 0 \quad \text{for } n = M, \cdots, N.$$

Then X_n is nonsingular for $n = M-1, \cdots, N$, and by Lemma 3.1, W_n as defined by (3.5) satisfies

$$X'_n L[X]_n = X'_n R[W]_n X_n = 0, \qquad n = M, \cdots, N.$$

Therefore $R[W]_n = 0$ for $n = M, \cdots, N$, from which (3.3) follows. Furthermore, since $X'_{n-1}C_{n-1}X_n > 0$ for $n = M, \cdots, N$, it follows from (3.11) that $W_n + C_{n-1}$ is positive definite for $n = M, \cdots, N$, and W_n is symmetric for $n = M, \cdots, N+1$. Thus (iv) implies (v).

(v)$\Rightarrow$(vi). Let W be a matrix sequence satisfying (v), and let η be a vector sequence defined for $n = M-1, \cdots, N$, satisfying $\eta_{M-1} = 0 = \eta_N$. We substitute A_n from (3.3) into (3.4) to obtain

$$J_2[\eta] = \sum_M^N (\Delta\eta_{n-1})' C_{n-1}\Delta\eta_{n-1} + \sum_M^N \eta'_n[W_{n+1} - W_n(W_n + C_{n-1})^{-1}C_{n-1}]\eta_n. \tag{3.18}$$

Since $\eta_{M-1} = 0 = \eta_N$, we have

$$\sum_M^N \eta'_n W_{n+1}\eta_n = \sum_{M-1}^{N-1} \eta'_n W_{n+1}\eta_n = \sum_M^N \eta'_{n-1} W_n \eta_{n-1}, \tag{3.19}$$

so the second sum on the right in (3.18) may be rewritten as

$$\sum_M^N [\eta'_{n-1} W_n \eta_{n-1} - \eta'_n W_n (W_n + C_{n-1})^{-1} C_{n-1}\eta_n], \tag{3.20}$$

while the first sum on the right in (3.18) may be expanded as

$$\sum_M^N [\eta'_n C_{n-1}\eta_n - \eta'_n C_{n-1}\eta_{n-1} - \eta'_{n-1}C_{n-1}\eta_n + \eta'_{n-1}C_{n-1}\eta_{n-1}]. \tag{3.21}$$

Adding (3.20) and (3.21) we obtain

$$\begin{aligned} J_2[\eta] = \sum_M^N \{ &\eta'_n[C_{n-1} - W_n(W_n + C_{n-1})^{-1}C_{n-1}]\eta_n \\ &+ \eta'_{n-1}(W_n + C_{n-1})\eta_{n-1} - \eta'_n C_{n-1}\eta_{n-1} - \eta'_{n-1}C_{n-1}\eta_n\}. \end{aligned} \tag{3.22}$$

Now

$$\begin{aligned} C_{n-1} - W_n(W_n + C_{n-1})^{-1}C_{n-1} &= [I - W_n(W_n + C_{n-1})^{-1}]C_{n-1} \\ &= [(W_n + C_{n-1}) - W_n](W_n + C_{n-1})^{-1}C_{n-1} \\ &= C_{n-1}(W_n + C_{n-1})^{-1}C_{n-1}, \end{aligned}$$

so (3.22) becomes

$$J_2[\eta]=\sum_{M}^{N}\{\eta_n' C_{n-1}(W_n+C_{n-1})^{-1}C_{n-1}\eta_n+\eta_{n-1}'(W_n+C_{n-1})\eta_{n-1} - \eta_n' C_{n-1}\eta_{n-1}-\eta_{n-1}' C_{n-1}\eta_n\}. \tag{3.23}$$

By use of the symmetry of W_n+C_{n-1}, this may be written in "completed-square" form as

$$J_2[\eta]=\sum_{M}^{N}\phi_n'(W_n+C_{n-1})^{-1}\phi_n \tag{3.24}$$

where

$$\phi_n=C_{n-1}\eta_n-(W_n+C_{n-1})\eta_{n-1}.$$

Thus $J_2[\eta]\geqq 0$. Furthermore, if $J_2[\eta]=0$ for some sequence η with $\eta_{M-1}=0=\eta_N$, then $\phi_n=0$, $n=M,\cdots,N$, i.e.,

$$C_{n-1}\eta_n=(W_n+C_{n-1})\eta_{n-1},\qquad n=M,\cdots,N.$$

Since $\eta_{M-1}=0$ and C_{n-1} is nonsingular, this implies that $\eta_n=0$ for $n=M,\cdots,N$. Thus $J_2[\eta]$ is positive definite, as claimed.

(vi)$\Rightarrow$(i). Assume $J_2[\eta]$ is positive definite on the given class of sequences η, and let u be a real vector solution of $l[u]=0$ with

$$u_{M-1}'C_{M-1}u_M\leqq 0 \quad\text{and}\quad u_M\neq 0. \tag{3.25}$$

For $p=M-1$ and $M-1<q<N$, define η as in Lemma 3.3, i.e., $\eta_n=u_n$ for $p+1\leqq n\leqq q$ and $\eta_n=0$ otherwise. Then $\eta_M=u_M\neq 0$ and $\eta_{M-1}=0=\eta_N$. Thus $J_2[\eta]>0$ and by Lemma 3.3,

$$\begin{aligned}0<J_2[\eta]&=u_{M-1}'C_{M-1}u_M+u_q'C_qu_{q+1}\\&\leqq u_q'C_qu_{q+1}\quad\text{for } q=M,\cdots,N-1.\end{aligned} \tag{3.26}$$

Thus (vi) implies (i).

(i)$\Leftrightarrow$(ii). Assume that (i) holds. Let U be the real matrix solution of $L[U]=0$ satisfying

$$U_{M-1}=0,\qquad U_M=I.$$

By Lemma 2.1, U is prepared, since $\{U,U\}_M=0$. For every nonzero constant vector z, the solution u of $l[u]=0$ defined by $u_n=U_nz$ satisfies

$$u_{M-1}'C_{M-1}u_M=z'U_{M-1}'C_{M-1}U_Mz=0,$$

and $u_M=U_Mz=z\neq 0$. Hence, by (i), $u_n'C_nu_{n+1}>0$ for $n=M,\cdots,N-1$, so

$$U_n'C_nU_{n+1}>0\quad\text{for } n=M,\cdots,N-1,$$

i.e.,

$$U_{n-1}'C_{n-1}U_n>0\quad\text{for } n=M+1,\cdots,N.$$

Thus U satisfies condition (iv) on the interval $n=M+1,\cdots,N$. Since we have already proved that (iv) implies (vi), it follows that the quadratic form defined by

$$J_2^*[\eta]=\sum_{M+1}^{N}[(\Delta\eta_{n-1})'C_{n-1}\Delta\eta_{n-1}+\eta_n'A_n\eta_n] \tag{3.27}$$

is positive definite on the class of vector sequences η which satisfy $\eta_M=0=\eta_N$.

Now let v be a real vector sequence of (1.1) satisfying the hypothesis of condition (ii), i.e.,

$$v'_{N-1}C_{N-1}v_N \leqq 0 \quad \text{and} \quad v_{N-1} \neq 0.$$

Application of Lemma 3.3 with $M-1$ replaced by M gives

$$0 < J_2^*[\eta] = v'_{p+1}C_pv_p + v'_qC_qv_{q+1}, \qquad M \leqq p < q < N,$$

for η defined by

$$\eta_n = \begin{cases} v_n, & p+1 \leqq n \leqq q, \\ 0, & \text{otherwise.} \end{cases}$$

In particular, for $q = N-1$, we have

$$0 < v'_{p+1}C_pv_p + v'_{N-1}C_{N-1}v_N \leqq v'_{p+1}C_pv_p \tag{3.28}$$

for $M \leqq p < N-1$, i.e., for $p = M, \cdots, N-2$. By Note 2.1, $v'_{p+1}C_pv_p = v'_pC_pv_{p+1}$ for all p, so (3.28) gives us

$$0 < v'_pC_pv_{p+1}, \qquad p = M, \cdots, N-2. \tag{3.29}$$

Finally we must show that (3.29) also holds for $p = M-1$. But, if not, then v is a real solution of (1.1) with

$$v'_{M-1}C_{M-1}v_M \leqq 0$$

and, from (3.29), $v_M \neq 0$, so by condition (i), $v'_{N-1}C_{N-1}v_N > 0$, contrary to our assumptions about v. Hence (i) implies (ii).

The proof that (ii) implies (i) is the dual of the preceding argument. We omit the details.

(ii)$\Rightarrow$(iii). Assume (ii). Then (i) also holds, and we actually prove here that (i) implies (iii). We must show that the interval $[M-1, N]$ contains no pair of conjugate intervals. By (i), $[M-1, M)$ is not conjugate to any other interval $[q, q+1)$ in $[M-1, N]$.

Suppose that two intervals $[p, p+1)$ and $[q, q+1)$ are conjugate for some p and q with $M \leqq p < q \leqq N-1$. Then (1.1) has a real vector solution y satisfying

$$y'_pC_py_{p+1} \leqq 0 \quad \text{and} \quad y'_qC_qy_{q+1} \leqq 0, \tag{3.30}$$

with $y_n \not\equiv 0$ on $[p+1, q]$ (which may be a single point, if $p+1 = q$). Let

$$\eta_n = \begin{cases} y_n, & p+1 \leqq n \leqq q, \\ 0 & \text{otherwise.} \end{cases}$$

Then by Lemma 3.3 and condition (3.30), $J_2^*[\eta]$ as defined in (3.27) satisfies

$$J_2^*[\eta] = y'_pC_py_{p+1} + y'_qC_qy_{q+1} \leqq 0. \tag{3.31}$$

But, as in the preceding proof, $J_2^*[\eta] > 0$, a contradiction. Thus (i), and hence (ii), implies (iii).

(iii)$\Rightarrow$(iv). Let U be the matrix solution of (1.2) satisfying

$$U_{M-1} = 0, \qquad U_M = I. \tag{3.32}$$

Then, as in the proof that (i) implies (ii), U is a prepared solution, and for every nonzero constant vector z, the solution $u_n = U_nz$ of (1.1) satisfies

$$u'_{M-1}C_{M-1}u_M = 0, \qquad u_M \neq 0.$$

Hence, by assumption (iii), there is no interval $[p, p+1)$ in $[M, N]$ conjugate to the interval $[M-1, M)$, so

$$u_n' C_n u_{n+1} > 0, \qquad n = M, \cdots, N-1. \tag{3.33}$$

Since $u_n' C_n u_{n+1} = z' U_n' C_n U_{n+1} z$, it follows that

$$U_n' C_n U_{n+1} > 0, \qquad n = M, \cdots, N-1; \tag{3.34}$$

hence U_n is nonsingular for $n = M, \cdots, N$.

Let V be the solution of (1.2) with

$$V_N = 0, \qquad V_{N-1} = (U_N' C_{N-1})^{-1}. \tag{3.35}$$

Since disconjugacy on $[M-1, N]$ implies (i), and (i) implies (ii), then by an argument similar to that leading to (3.34) we obtain

$$V_n' C_n V_{n+1} > 0, \qquad n = M-1, \cdots, N-2. \tag{3.36}$$

Also,

$$\begin{aligned} \{U_n, V_n\} \equiv \{U_N, V_N\} &= U_{N-1}' C_{N-1} V_N - U_N' C_{N-1} V_{N-1} \\ &= -U_N' C_{N-1} (U_N' C_{N-1})^{-1} = -I; \end{aligned} \tag{3.37}$$

hence by (2.5), for all n for which U_n and V_n are defined,

$$\{V_n, U_n\} = -\{U_n, V_n\}' = I. \tag{3.38}$$

Now let X be the solution of (1.2) defined by $X_n = U_n + V_n$. Then X is a prepared solution, since

$$\begin{aligned} \{X_n, X_n\} &= \{U_n + V_n, U_n + V_n\} = \{U_n, U_n\} + \{U_n, V_n\} + \{V_n, U_n\} + \{V_n, V_n\} \\ &= 0 - I + I + 0 = 0. \end{aligned}$$

We must show that

$$X_{n-1}' C_{n-1} X_n > 0 \quad \text{for } n = M, \cdots, N. \tag{3.39}$$

First, for $n = M$, we note that

$$-I = \{U_M, V_M\} = U_{M-1}' C_{M-1} V_M - U_M' C_{M-1} V_{M-1} = -C_{M-1} V_{M-1}$$

so

$$C_{M-1} V_{M-1} = I. \tag{3.40}$$

Then

$$\begin{aligned} X_{M-1}' C_{M-1} X_M &= (U_{M-1}' + V_{M-1}') C_{M-1} (U_M + V_M) \\ &= V_{M-1}' C_{M-1} (U_M + V_M) = V_{M-1}' C_{M-1} + V_{M-1}' C_{M-1} V_M \\ &= I + V_{M-1}' C_{M-1} V_M > 0 \end{aligned}$$

by (3.36), so (3.39) holds for $n = M$.

For $n = N$, using (3.34) and (3.35) we obtain

$$\begin{aligned} X_{N-1}' C_{N-1} X_N &= (U_{N-1}' + V_{N-1}') C_{N-1} (U_N + V_N) \\ &= U_{N-1}' C_{N-1} U_N + V_{N-1}' C_{N-1} U_N \\ &= U_{N-1}' C_{N-1} U_N + (C_{N-1} U_N)^{-1} C_{N-1} U_N \\ &= U_{N-1}' C_{N-1} U_N + I > 0, \end{aligned}$$

so (3.39) holds for $n = N$.

Finally, suppose that $M<k<N$, and let α be an arbitrary unit vector. We must show that

$$\alpha' X'_{k-1} C_{k-1} X_k \alpha > 0. \tag{3.41}$$

Let

$$\eta_n = \begin{cases} U_n\alpha, & M-1 \leqq n \leqq k-1, \\ -V_n\alpha, & k \leqq n \leqq N+1. \end{cases} \tag{3.42}$$

Then by definition of U and V, η satisfies

$$\begin{aligned} &\eta_{M-1} = U_{M-1}\alpha = 0, \qquad \eta_N = -V_N\alpha = 0, \\ &\eta_M = U_M\alpha = \alpha \neq 0, \qquad \Delta\eta_{M-1} = \eta_M - \eta_{M-1} = \alpha. \end{aligned} \tag{3.43}$$

We will show that $J_2[\eta]$ as defined in (vi) satisfies

$$0 \leqq J_2[\eta] = \alpha' X'_{k-1} C_{k-1} X_k \alpha + \alpha'\{U, V\}\alpha = \alpha' X'_{k-1} C_{k-1} X_k \alpha - 1, \tag{3.44}$$

from which (3.41) follows.

Consider the prepared solution U of (1.2) determined by the initial conditions (3.32). This solution satisfies (3.34), which we rewrite as

$$U'_{n-1} C_{n-1} U_n > 0, \qquad n = M+1, \cdots, N. \tag{3.45}$$

Thus U satisfies condition (iv) on the interval $[M+1, N]$. Since (iv) implies (v) there exists a symmetric sequence W defined by

$$W_n = (C_{n-1}\Delta U_{n-1})U_{n-1}^{-1}, \qquad n = M+1, \cdots, N+1, \tag{3.46}$$

with $W_n + C_{n-1} > 0$ for $n = M+1, \cdots, N$, and

$$W_{n+1} - A_n = W_n(W_n + C_{n-1})^{-1} C_{n-1}, \qquad n = M+1, \cdots, N. \tag{3.47}$$

Since our sequence η defined by (3.42) fails to have $\eta_M = 0$, the argument used above in (v)$\Rightarrow$(vi) to show $J_2[\eta]$ positive definite does not hold here, and we must treat $J_2[\eta]$ somewhat differently in this case. Using (3.43) we obtain

$$J_2[\eta] = \alpha' C_{M-1}\alpha + \alpha' A_M \alpha + \sum_{M+1}^{N} [(\Delta\eta_{n-1})' C_{n-1} \Delta\eta_{n-1} + \eta'_n A_n \eta_n],$$

so, by substitution for A_n from (3.47),

$$\begin{aligned} J_2[\eta] = \alpha' C_{M-1}\alpha + \alpha' A_M \alpha &+ \sum_{M+1}^{N} (\Delta\eta_{n-1})' C_{n-1} \Delta\eta_{n-1} \\ &+ \sum_{M+1}^{N} [\eta'_n(W_{n+1} - W_n(W_n + C_{n-1})^{-1} C_{n-1})\eta_n]. \end{aligned} \tag{3.48}$$

Now

$$\sum_{M+1}^{N} \eta'_n W_{n+1} \eta_n = \sum_{M}^{N-1} \eta'_n W_{n+1} \eta_n + \eta'_N W_{N+1} \eta_N - \eta'_M W_{M+1} \eta_M.$$

By a shift of indices in the sum on the right, and use of $\eta_N = 0$, this becomes

$$\sum_{M+1}^{N} \eta'_n W_{n+1} \eta_n = \sum_{M+1}^{N} \eta'_{n-1} W_n \eta_{n-1} - \eta'_M W_{M+1} \eta_M.$$

From (3.43) and (3.46) this yields

$$\sum_{M+1}^{N} \eta'_n W_{n+1} \eta_n = \sum_{M+1}^{N} \eta'_{n-1} W_n \eta_{n-1} - \alpha'(C_M \Delta U_M) U_M^{-1} \alpha. \tag{3.49}$$

Using (3.49), we write the second sum on the right in (3.48) as

$$(3.50)\quad \sum_{M+1}^{N} (\eta'_{n-1} W_n \eta_{n-1} - \eta'_n W_n (W_n + C_{n-1})^{-1} C_{n-1} \eta_n) - \alpha'(C_M \Delta U_M) U_M^{-1} \alpha$$

while the first sum on the right in (3.48) can be expanded as

$$(3.51)\quad \sum_{M+1}^{N} (\eta'_n C_{n-1} \eta_n - \eta'_n C_{n-1} \eta_{n-1} - \eta'_{n-1} C_{n-1} \eta_n + \eta'_{n-1} C_{n-1} \eta_{n-1}).$$

Substituting (3.50) and (3.51) into (3.48) gives us

$$\begin{aligned} J_2[\eta] = {} & \alpha'[C_{M-1} + A_M - (C_M \Delta U_M) U_M^{-1}]\alpha \\ & + \sum_{M+1}^{N} \{\eta'_n [C_{n-1} - W_n (W_n + C_{n-1})^{-1} C_{n-1}]\eta_n \\ & \qquad + \eta'_{n-1}(W_n + C_{n-1})\eta_{n-1} - \eta'_n C_{n-1} \eta_{n-1} - \eta'_{n-1} C_{n-1} \eta_n\}. \end{aligned} \tag{3.52}$$

The same steps that led from (3.22) to (3.24) now yield

$$(3.53)\quad J_2[\eta] = \alpha'[C_{M-1} + A_M - (C_M \Delta U_M) U_M^{-1}]\alpha + \sum_{M+1}^{N} \phi'_n (W_n + C_{n-1})^{-1} \phi_n,$$

where

$$\phi_n = C_{n-1}\eta_n - (W_n + C_{n-1})\eta_{n-1}.$$

We will show that the first term on the right in (3.53) equals zero. Since $U_M = I$, we have

$$(3.54)\quad C_{M-1} + A_M - (C_M \Delta U_M) U_M^{-1} = C_{M-1} + A_M - (C_M U_{M+1} - C_M) = B_M - C_M U_{M+1},$$

where, as in (1.3) and (1.4),

$$B_M = C_M + C_{M-1} + A_M.$$

Now $L[U] = 0$, so from (1.4) we know

$$-C_M U_{M+1} - C_{M-1} U_{M-1} + B_M U_M = 0.$$

Since $U_{M-1} = 0$ and $U_M = I$, this becomes

$$-C_M U_{M+1} + B_M = 0.$$

Thus it follows from (3.54) that the first term on the right in (3.53) equals zero, so

$$J_2[\eta] = \sum_{M+1}^{N} \phi'_n (W_n + C_{n-1})^{-1} \phi_n \geqq 0.$$

If $J_2[\eta] = 0$, then $\phi_n = 0$, $n = M+1, \cdots, N$, i.e.,

$$(3.55)\quad C_{n-1}\eta_n = (W_n + C_{n-1})\eta_{n-1}, \qquad n = M+1, \cdots, N.$$

Since $\eta_N = 0$ and $W_n + C_{n-1}$ is nonsingular for $n = M+1, \cdots, N$, (3.55) implies $\eta_M = 0$, contrary to (3.43). Thus $J_2[\eta] > 0$. Let $u_n = U_n \alpha$ and $v_n = V_n \alpha$, $n = M-1, \cdots, N+1$,

so that (3.42) may be written as

$$\eta_n = \begin{cases} u_n, & M-1 \leqq n \leqq k-1, \\ -v_n, & k \leqq n \leqq N+1. \end{cases}$$

Thus

$$\Delta\eta_n = \begin{cases} \Delta u_n, & M-1 \leqq u \leqq k-2, \\ -v_k - u_{k-1}, & n = k-1, \\ -\Delta v_n, & k \leqq n \leqq N, \end{cases}$$

and $J_2[\eta]$ becomes

$$\begin{aligned} J_2[\eta] = &\sum_{M}^{k-1} [(\Delta u_{n-1})' C_{n-1} \Delta u_{n-1} + u_n' A_n u_n] + (-v_k' - u_{k-1}') C_{k-1}(-v_k - u_{k-1}) + v_k' A_k v_k \\ &+ \sum_{k+1}^{N} [(\Delta v_{n-1})' C_{n-1} \Delta v_{n-1} + v_n' A_n v_n]. \end{aligned} \tag{3.56}$$

Application of Lemma 3.2 to the two sums on the right in (3.56) yields

$$\begin{aligned} J_2[\eta] = &u_{k-1}' C_{k-1} \Delta u_{k-1} - u_{M-1}' C_{M-1} \Delta u_{M-1} + \sum_{M}^{k-1} u_n' l[u]_n \\ &+ (v_k' C_{k-1} v_k + v_k' C_{k-1} u_{k-1} + u_{k-1}' C_{k-1} v_k + u_{k-1}' C_{k-1} u_{k-1} + v_k' A_k v_k) \\ &+ v_N' C_N \Delta v_N - v_k' C_k \Delta v_k + \sum_{k+1}^{N} v_n' l[v]_n. \end{aligned} \tag{3.57}$$

But u and v satisfy $l[u] = 0$ and $l[v] = 0$, and also $u_{M-1} = 0 = v_N$, so (3.57) reduces to

$$\begin{aligned} J_2[\eta] &= u_{k-1}' C_{k-1}(u_k - u_{k-1}) + v_k' C_{k-1} v_k + v_k' C_{k-1} u_{k-1} \\ &\quad + u_{k-1}' C_{k-1} v_k + u_{k-1}' C_{k-1} u_{k-1} + v_k' A_k v_k - v_k' C_k (v_{k+1} - v_k) \\ &= u_{k-1}' C_{k-1} u_k + u_{k-1}' C_{k-1} v_k + v_k' (B_k v_k + C_{k-1} u_{k-1} - C_k v_{k+1}). \end{aligned} \tag{3.58}$$

Since v is a solution of (1.3), this becomes

$$J_2[\eta] = u_{k-1}' C_{k-1} u_k + u_{k-1}' C_{k-1} v_k + v_k' C_{k-1} u_{k-1} + v_k' C_{k-1} v_{k-1}. \tag{3.59}$$

The last term in (3.59) equals $v_{k-1}' C_{k-1} v_k$, by Note 2.1, and (3.59) may thus be rewritten as

$$J_2[\eta] = \alpha'(U_{k-1}' C_{k-1} U_k + U_{k-1}' C_{k-1} V_k + V_k' C_{k-1} U_{k-1} + V_{k-1}' C_{k-1} V_k)\alpha. \tag{3.60}$$

Now for the solution $X = U + V$ of (1.2), we have

$$X_{k-1}' C_{k-1} X_k = U_{k-1}' C_{k-1} U_k + U_{k-1}' C_{k-1} V_k + V_{k-1}' C_{k-1} U_k + V_{k-1}' C_{k-1} V_k. \tag{3.61}$$

Also, from (3.38),

$$\{V_k, U_k\} = V_{k-1}' C_{k-1} U_k - V_k' C_{k-1} U_{k-1} = I, \tag{3.62}$$

and combining (3.60), (3.61), and (3.62), we see that

$$\begin{aligned} J_2[\eta] &= \alpha'(X_{k-1}' C_{k-1} X_k - \{V_k, U_k\})\alpha \\ &= \alpha' X_{k-1}' C_{k-1} X_k \alpha - \alpha'\alpha \\ &= \alpha' X_{k-1}' C_{k-1} X_k \alpha - 1, \end{aligned}$$

since α was an arbitrary unit vector. Since $J_2[\eta]>0$, it follows that $X'_{k-1}C_{k-1}X_k$ is positive definite, which completes the proof of the theorem.

Note added in proof. A referee has raised the question of conditions equivalent to disconjugacy on $[M-1,\infty)$. The following result gives a set of related, but not equivalent, conditions.

COROLLARY 3.4. *Assume hypotheses* (3.1)-(3.2). *The following conditions are related by* $(\alpha)\Rightarrow(\beta)\Rightarrow(\gamma)\Rightarrow(\delta)$.

(α) *Equation* (1.1) *is disconjugate on* $[M-1,\infty)$.

(β) *The matrix solution* U_n *defined by* $U_{M-1}=0$, $U_M=I$ *has* $U'_nC_nU_{n+1}>0$, $n=M,\cdots$.

(γ) *There exists a sequence* W *of symmetric matrices with* $W_n+C_{n-1}>0$ *for* $n=M+1,\cdots$, *satisfying* (3.3) *for* $n=M+1,\cdots$.

(δ) *Equation* (1.1) *is disconjugate on* $[M,\infty)$.

REFERENCES

[1] C. D. AHLBRANDT AND J. W. HOOKER, *Riccati transformations and principal solutions of discrete linear systems*, in Proc. 1984 Workshop Spectral Theory of Sturm-Liouville Differential Operators, H. G. Kaper and A. Zettl, eds., ANL-84-87, Argonne National Laboratories, Argonne, IL, 1984.

[2] ———, *Disconjugacy criteria for second order linear difference equations*, in Proc. International Conference on Qualitative Theory of Differential Equations, Edmonton, Alberta, Canada, 1984.

[3] ———, *A variational view of nonoscillation theory for linear difference equations*, in Proc. Thirteenth Midwest Differential Equations Conference, J. L. Henderson, ed., University of Missouri-Rolla, Rolla, MO, 1985.

[4] F. W. ATKINSON, *Discrete and Continuous Boundary Problems*, Academic Press, New York, 1964.

[5] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Mathematics 220, Springer-Verlag, New York, 1971.

[6] T. FORT, *Finite Differences and Difference Equations in the Real Domain*, Oxford University Press, London, 1948.

[7] P. HARTMAN, *Difference equations: disconjugacy, principal solutions, Green's functions, complete monotonicity*, Trans. Amer. Math. Soc., 246 (1978), pp. 1-30.

[8] ———, *Self-adjoint, non-oscillatory systems of ordinary, second order, linear differential equations*, Duke Math. J., 24 (1957), pp. 25-36.

[9] D. B. HINTON AND R. T. LEWIS, *Spectral analysis of second order difference equations*, J. Math. Anal. Appl., 63 (1978), pp. 421-438.

[10] ———, *Oscillation theory for generalized second-order differential equations*, Rocky Mountain J. Math., 10 (1980), pp. 751-766.

[11] J. W. HOOKER AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations*, J. Math. Anal. Appl., 82 (1981), pp. 451-462.

[12] J. W. HOOKER, M. K. KWONG, AND W. T. PATULA, *Riccati type transformations for second-order linear difference equations* II, J. Math. Anal. Appl., 107 (1985), pp. 182-196.

[13] ———, *Oscillatory second order linear difference equations and Riccati equations*, SIAM J. Math. Anal., 18 (1987), pp. 54-63.

[14] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Trans. ASME, Ser. D, J. Basic Engrg., 82 (1960), pp. 35-45.

[15] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[16] A. B. MINGARELLI, *Volterra-Stieltjes Integral Equations and Generalized Ordinary Differential Expressions*, Lecture Notes in Mathematics 989, Springer-Verlag, New York, 1983.

[17] W. T. PATULA, *Growth, oscillation, and comparison theorems for second order linear difference equations*, SIAM J. Math. Anal., 10 (1979), pp. 1272-1279.

[18] A. C. PETERSON, *Boundary value problems for an nth order linear difference equation*, SIAM J. Math. Anal., 15 (1984), pp. 124-132.

[19] W. T. REID, *Oscillation criteria for linear differential systems with complex coefficients*, Pacific J. Math., 6 (1956), pp. 733-751.

[20] ———, *Riccati matrix differential equations and non-oscillation criteria for associated linear differential systems*, Pacific J. Math., 10 (1963), pp. 665-685.

[21] ———, *Ordinary Differential Equations*, John Wiley, New York, 1971.

[22] W. T. REID, *Riccati Differential Equations*, Mathematics in Science and Engineering, 86, Academic Press, New York, 1972.

[23] ———, *A criterion of oscillation for generalized differential equations*, Rocky Mountain J. Math., 7 (1977), pp. 799–806.

[24] D. R. VAUGHAN, *A nonrecursive algebraic solution for the discrete Riccati equation*, IEEE Trans. Automat. Control, 15 (1970), pp. 597–599.

[25] N. WATANABE, *Note on the Kalman filter with estimated parameters*, J. Time Ser. Anal., 6 (1985), pp. 269–278.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

016

SAMPLING BANDLIMITED FUNCTIONS OF POLYNOMIAL GROWTH*

GILBERT G. WALTER†

Abstract. Two new versions of the sampling theorem extended to functions whose Fourier transform is a generalized function are given. One involves a correction by means of an arbitrary polynomial and the other involves (C, α)-summability. The best approximation to nonbandlimited functions of polynomial growth by functions whose transform has compact support in the Sobolev norm is found.

Key words. sampling theorem, bandlimited signal, generalized functions, Sobolev spaces, best approximation

AMS(MOS) subject classifications. 41A05, 40G05

1. Introduction. The sampling theorem for bandlimited signals is the name usually given to the formula

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{\sin \sigma(t-nT)}{\sigma(t-nT)} \tag{1.1}$$

where $T = \pi/\sigma$ is the sampling period. "Bandlimited" means the Fourier transform $F(w)$ of $f(t)$ is zero for $|w| > \sigma$.

This theorem is well established for $F(w)$ which are L^2 functions, and goes back to Whittaker [8], but was developed and exploited by Shannon [7] much later. Subsequently it was generalized in a number of different directions [2], in particular to f such that $F(w)$ is a generalized function with compact support [1], [3]-[5]. In this work we shall extend it further in this direction.

Such functions f belong to L^2_r the space of functions on $\mathbb{R}^1$ square integrable with respect to the measure $dm_r = (1+t^2)^{-r}\,dt$, $r = 0, 1, \cdots$. Following [3], we denote by $B_r(\sigma)$ those $f \in L^2_r$ which are bandlimited to $[-\sigma, \sigma]$. These functions are also, by the Schwartz-Paley-Wiener Theorem, entire functions whose restriction to the real axis is of polynomial growth. Hence we shall refer to them as *bandlimited functions of polynomial growth.*

Several versions of the sampling theorem for bandlimited functions of polynomial growth have appeared. The first were due to Campbell [1] and to Pfaffelhuber [5]. The former deals with signals $f(t)$ whose Fourier transform $F(w)$ has compact support in the *interior* of the interval $(-\sigma, \sigma)$, where $S(t) = S_\varepsilon(t)$ is an appropriate smoothing function used to obtain convergence in (1.1).

Pfaffelhuber showed that the generalized sampling theorem can be extended to the case where $F(w)$ has support in the closed interval $[-\sigma, \sigma]$. He replaced (1.1) with the expression

$$f(t) = q_N(t)\cos \sigma t + \sum_{n=-\infty}^{\infty} (f(nT) - q_N(nT)(-1)^n)\left(\frac{t}{nT}\right)^N \frac{\sin \sigma(t-nT)}{\sigma(t-nT)}, \tag{1.2}$$

where N is the order of the generalized function $F(w)$ and $q_N(t)$ is an appropriate polynomial of degree $\leqq N-1$.

Lee [4] replaced the $\cos \sigma t$ by e^{iat}, $|a| \leqq \sigma$ in (1.2), and thus was able to replace $q_N(t)$ by the Taylor polynomial of $f(t)$ for $a = 0$.

* Received by the editors September 18, 1986; accepted for publication (in revised form) October 1, 1987. This research was supported in part by National Science Foundation grant DCR-8504620.

† Department of Mathematical Sciences, University of Wisconsin, Milwaukee, Wisconsin 53201.

Hoskins and De Sousa Pinto [3] generalized both results by replacing $\cos \sigma t$ and $e^{i\alpha t}$ by an arbitrary bandlimited function $\eta(t)$, $\eta(0) \neq 0$, whose Fourier transform is a measure with support in $[-\sigma, \sigma]$.

In this work we shall first generalize Lee's result in a different direction, to polynomials other than the Taylor polynomials. We then show that for certain such polynomials the sampling expansion converges in L_r^2 and obtain the best approximation to $f \in L_r^2$ by a function $g \in B_r(\sigma)$. Finally we give a version of the sampling theorem for bandlimited functions of polynomial growth having the same form as (1.1) except that convergence is replaced by (C, α)-summability.

2. Sampling expansions with polynomials correction. In this section we prove a sampling theorem for bandlimited functions of polynomial growth which involves an arbitrary polynomial of the same degree as the order of $F(w)$, the Fourier transform of f. We shall use the convention that F is of order r if $f \in L_r^2$, that is, that F is given by an rth order differential operator applied to an L^2 function. We shall consider polynomials of the form

$$P(t) = \prod_{j=1}^{k} (t - a_j)^{m_j}, \qquad \sum_{j=1}^{k} m_j = N, \tag{2.1}$$

a_j possibly complex.

LEMMA 2.1. *Let $f(t)$ be a signal in $B_N(\sigma)$ such that $f^{(i)}(a_j) = 0$, $i = 0, 1, \cdots, m_j - 1$, $j = 1, 2, \cdots, k$. Then, with P given in* (2.1),

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{P(t)}{P(nT)} \frac{\sin \sigma(t - nT)}{\sigma(t - nT)} \tag{2.2}$$

where the convergence is uniform on bounded subsets of the complex t-plane.

Proof. We first note that $B_N(\sigma)$ is closed under translation. Hence $f(t)/P(t)$ may be obtained by repeated division by t and translation. Accordingly let $\psi(t) \in B_K(\sigma)$, $1 \leqq K \leqq N$, $\psi(0) = 0$, and $\varphi(t) = \psi(t)/t$.

By the Schwartz-Paley-Wiener Theorem, ψ is an entire function of exponential type $\leqq \sigma$, and since $B_K(\sigma) \subset L_K^2$, $\psi(t)/(t^2+1)^{K/2} \in L^2(\mathbb{R}^1)$. Since the exponential type of an entire function depends only on its Taylor coefficients, it follows that $\varphi(t)$ is also entire of exponential type $\leqq \sigma$. Moreover, $\varphi(t)/(t^2+1)^{(K-1)/2} \in L^2(\mathbb{R}^1)$, and hence $\varphi(t) \in B_{K-1}(\sigma)$.

We return now to our original $f(t)$ and conclude that $f(t)/P(t) \in B_0(\sigma) \subset L^2(R^1)$. By applying the sampling theorem for L^2 functions to $h(t) = f(t)/P(t)$, we obtain (2.2).

Remark 2.1. Among the interesting cases for (2.1) are the following:

(i) $P(t) = (t - a)^N$;

(ii) All a_j distinct, $P(t) = \prod_{j=1}^{N} (t - a_j)$;

(iii) $P(t) = (t^2+1)^M$, $N = 2M$.

Each gives us a different application.

The first gives us Lee's result with Taylor's polynomial but at an arbitrary point instead of zero. Case (ii) is an interpolation result that is particularly useful when the values of f are known only at a discrete set. We state it as follows.

COROLLARY 2.2. *Let $P(t) = \prod_{j=1}^{N} (t - a_j)$ where all a_j are distinct, $f(t) \in B_N(\sigma)$; let $L_N(t)$ be the polynomial that interpolates f at the points $(a_1, \cdots, a_N)$. Then*

$$f(t) - L_N(t) = \sum_{n=-\infty}^{\infty} (f(nT) - L_N(nT)) \frac{P(t)}{P(nT)} \frac{\sin \sigma(t - nT)}{\sigma(t - nT)}.$$

Case (iii) is useful when we want to obtain convergence in the sense of L_r^2.

3. Convergence and approximation in L_r^2. In this section we deal with convergence in L_r^2 of the sampling expansion of bandlimited signals of polynomial growth. We use a form of $P(t)$, the polynomial in (2.2), which is particularly simple and which enables us to remove the restrictions on $f(t)$. Since $B_r(\sigma)$ is a proper subspace of L_r^2, we may consider best approximations in the former to an arbitrary element in L_r^2.

We should note that the Fourier transform maps L_r^2 into the Sobolev space H^{-r} isometrically. See Rudin [6] for details. This latter space is composed of distributions of the form $F=(1-D^2)^N G$ for $r=2N$ where $G\in L^2(\mathbb{R}^1)$.

PROPOSITION 3.1. *Let N be a positive integer and let $f\in B_{2N}(\sigma)$; then there exists a polynomial $P(t)$ of degree $2N-1$ such that*

$$f(t)-P(t)\sin\sigma t=\sum_{n=-\infty}^{\infty} f(nT)\left[\frac{t^2+1}{n^2T^2+1}\right]^N \frac{\sin\sigma(t-nT)}{\sigma(t-nT)} \tag{3.1}$$

where the series converges in the sense of L_{2N}^2 and uniformly on compact sets.

Proof. There is a polynomial $d(t)$ of degree $\leqq 2N-1$ such that $f(t)-d(t)$ has N-fold zeros at $t=\pm i$. Then by Lemma 2.1,

$$f(t)-d(t)=\sum_{n=-\infty}^{\infty}[f(nT)-d(nT)]\left[\frac{t^2+1}{n^2T^2+1}\right]^N \frac{\sin\sigma(t-nT)}{\sigma(t-nT)}. \tag{3.2}$$

This series is easily seen to converge in the sense of L_{2N}^2 since $s_n(t)=\sin\sigma(t-nT)/\sigma(t-nT)$ is an orthonormal sequence. The Fourier transform of $d(t)$ is of the form

$$D(w)=\sum_{j=0}^{2N-1} a_j\delta^{(j)}(w), \tag{3.3}$$

and hence has a Fourier series on $(-\sigma,\sigma)$

$$D(w)=\sum c_n e^{-iwnT}, \qquad |w|<\sigma, \tag{3.4}$$

with coefficients $c_n=d(nT)=O(n^{2N-1})$, which converges to the periodic extension $D^*(w)$ for real w. But $D^*(w)=(1-D_w^2)^N G^*(w)$ where D_w is the differentiation operator and

$$G^*(w)=\sum\frac{c_n}{(1+n^2T^2)^N}e^{-iwnT}\in L^2(-\sigma,\sigma). \tag{3.5}$$

Let $G(w)=G^*(w)\chi(w)$ where χ is the characteristic function of $[-\sigma,\sigma]$. Then G is in $L^2(\mathbb{R}^1)$ and its inverse Fourier transform $g(t)$ is given by

$$g(t)=\sum\frac{d(nT)}{(1+n^2T^2)^N}\frac{\sin\sigma(t-nT)}{\sigma(t-nT)}, \tag{3.6}$$

where $g\in L^2(R^1)$. Therefore $(1+t^2)^N g(t)\in L_{2N}^2$ and the series

$$\sum d(nT)\left(\frac{1+t^2}{1+n^2T^2}\right)^N\frac{\sin\sigma(t-nT)}{\sigma(T-nT)} \tag{3.7}$$

converges to $(1+t^2)^N g(t)$ in the sense of L_{2N}^2. Since $D^*(w)$ has support on 0, $\pm 2\sigma$, $\pm 4\sigma,\cdots$, it is locally equal to zero on open intervals excluding these points. Hence G^* is a classical solution to $(1-D^2)^N Y=0$ on $(\sigma/2,3\sigma/2)$ and on $(-3\sigma/2,-\sigma/2)$ and is locally a C^∞ function at $w=\pm\sigma$. Since it is also periodic, it follows by repeated integration by parts that

$$d(t)=\frac{1}{2\sigma}\int_{-\sigma}^{\sigma}(1-D^2)^N G(w)\,e^{iwt}\,dw=(1+t^2)^N g(t)+P(t)\sin\sigma t. \tag{3.8}$$

The result for best approximation in L^2_{2p} in the proper norm is somewhat more complex than in L^2.

PROPOSITION 3.2. *Let $f \in L^2_{2p}$, $g(t) = f(t)/(t^2+1)^p$; then the best approximation to f among $y \in B_{2p}(\sigma)$ is obtained for*

$$y_0 = (t^2+1)^p \left(g(t) * \frac{\sin \sigma t}{\sigma t} \right) + \sum_{j=0}^{p-1} a_j t^j e^{i\sigma t} + \sum_{j=0}^{p-1} b_j t^j e^{-i\sigma t} \tag{3.9}$$

where $a_0, \cdots, a_{p-1}, b_0, \cdots, b_{p-1}$ depend on f.

Proof. Let $F(w)$ and $G(w)$ be the Fourier transform of f and g, respectively. Then $F(w) = (1-D^2)^p G(w)$ and $G \in L^2(\mathbb{R}^1)$. We split up G into three mutually orthogonal functions as follows: first we let

$$G = G\chi + G(1-\chi)$$

where χ is the characteristic function of $[-\sigma, \sigma]$. We then take $G_1 = G\chi$ and G_2 to be the projection of $G(1-\chi)$ on to the subspace of $L^2(R^1)$ spanned by functions of the form

$$w^k e^{-w} H(w-\sigma), \quad w^k e^{w} H(-w-\sigma), \quad k = 0, 1, \cdots, p-1$$

where $H(w)$ is the Heaviside function. We take $G_3 = G(1-\chi) - G_2$ which is clearly orthogonal to G_2. Hence we have a decomposition

$$G = G_1 + G_2 + G_3$$

into mutually orthogonal functions.

We now let $y \in B_{2p}(\sigma)$. Its Fourier transform Y has a decomposition $Y = (1-D^2)^p(Z_1 + Z_2 + Z_3)$ and the norm of the difference is

$$\|f - y\|^2_{2p} = \|G_1 + G_2 + G_3 - Z_1 - Z_2 - Z_3\|^2. \tag{3.10}$$

But since $(1-D^2)^p(Z_2+Z_3) = 0$ on $(-\infty, -\sigma) \cup (\sigma, \infty)$, it follows that $Z_2 + Z_3$ is a polynomial of degree $\leqq p-1$ times e^{-w} on (σ, ∞), and hence $Z_3 = 0$ for $w > \sigma$. A similar result holds for $-\sigma > w$. Hence (3.10) is minimized by choosing $Z_1 = G_1$ and $Z_2 = G_2$, i.e., the inverse Fourier transform y_0 of

$$Y_0 = (1-D^2)^p (G_1 + G_2)$$

is the best approximation. It is

$$y_0(t) = (1+t^2)^p \mathcal{F}^{-1}(G\chi) + \mathcal{F}^{-1} \sum_{n=0}^{p-1} a_n \delta^{(n)}(w-\sigma) + \mathcal{F}^{-1} \sum_{n=0}^{p-1} b_n \delta^{(n)}(w+\sigma),$$

where the coefficients a_n and b_n involve the value of G_2 and its derivatives at $w = \pm\sigma$. This gives us the theorem.

4. (C, α)-summability. In this section we shall consider a version of (1.1) appropriate for bandlimited functions of polynomial growth which, however, involve (C, α) Cesaro summability instead of pointwise convergence. That is, we shall show that

$$f(t) = \lim_{N \to \infty} \sum_{n=-N}^{N} C^{\alpha}_{N,n} f(nT) \frac{\sin \sigma(t-nT)}{\sigma(t-nT)} \tag{4.1}$$

where

$$C^{\alpha}_{N,n} = A^{\alpha}_{N-|n|} / A^{\alpha}_{N}, \qquad A^{\alpha}_{k} = \binom{k+\alpha}{k}.$$

This is the same weight used in (C, α)-summability in Fourier series [10a, p. 77]. This result is closer to the original formulation than those considered in the previous sections.

We suppose that $f(t)$ is a bandlimited function of polynomial growth with $F(w)$ having support in $[-\sigma, \sigma]$. We shall need a characterization of such F in terms of piecewise continuous functions. It is similar to that in [3] and could be derived from it.

The space of generalized functions with compact support is denoted by $\mathscr{E}'$ $\mathscr{E}'$ is a subspace of S', the space of all tempered distributions, and thus each element is a finite order derivative of a continuous function of polynomial growth. We use this for our characterization.

LEMMA 4.1. *Let $F \in \mathscr{E}'$ with support in $[-\sigma, \sigma]$; then there exists a piecewise continuous function $G \in \mathscr{E}'$ with the same support as F, an integer p, and constants $c_0, \cdots, c_{p-1}$ such that*

$$F(w) = D^p G(w) + \sum_{i=0}^{p-1} c_i \delta^i(w). \tag{4.2}$$

Proof. We begin by observing that each $F \in S'$ has an antiderivative in S', $F^{(-1)}$. Since $F = 0$ on $(-\infty, -\sigma) \cup (\sigma, \infty)$, $F^{(-1)}$ is constant on these two intervals and we may assume without loss of generality that $F^{(-1)}(w) = 0$ on $(-\infty, -\sigma)$. Let $a_1 = F^{(-1)}(w)$ on (σ, ∞). Then

$$F^{(-1)}(w) - a_1 H(w),$$

where $H(w)$ is the unit step function, is zero on $(-\infty, -\sigma) \cup (\sigma, \infty)$ and hence belongs to $\mathscr{E}'$. We repeat this argument $(p-1)$ times to obtain an element

$$G(w) = F^{(-p)}(w) - a_1 \frac{w^{p-1}}{(p-1)!} H(w) - \frac{a_2 w^{p-2}}{(p-2)!} H(w) - \cdots - a_p H(w)$$

in $\mathscr{E}'$ with support in $[-\sigma, \sigma]$. But for p sufficiently large $F^{(-p)}$ is a continuous function, and hence $G(w)$ is continuous except at 0. The pth derivative of G is the expression given in (4.2).

We now turn to (C, α)-summability of the Fourier series of $F(w)$. We need the following.

LEMMA 4.2. *Let $\varphi_t(w) = e^{iwt} \chi_{[-\pi, \pi]}(w)$, $t \in \mathbb{R}^1$, and let φ_t^* be its periodic extension. Then the Fourier series of the pth derivative of φ_t^* is uniformly (C, α)-summable on closed subintervals of $(-\pi, \pi)$ for t on bounded sets when $\alpha > p$.*

Proof. The fact that $D^p \varphi_t^*(w)$ has a Fourier series which is (C, α)-summable at each $w \in (-\pi, \pi)$ follows from a classical theorem [10b, p. 59] since the pth derivative of φ_t^* exists at each such w. The same proof gives uniform convergence provided w is restricted to an interval $[-\pi + \varepsilon, \pi - \varepsilon]$ since the pth derivative of φ_t^* is uniformly continuous in this interval.

We now use these lemmas to obtain (C, α)-summability of the Fourier transform of $F \in \mathscr{E}'$.

THEOREM 4.3. *Let $F \in \mathscr{E}'$ with support in the interior of $(-\pi, \pi)$; let $f(t) = 1/2\pi \langle F, e^{-iwt} \rangle$ and $f_n^\alpha(t) = 1/2\pi \langle F, \sigma_{nt}^\alpha \rangle$ where $\sigma_{nt}^\alpha(w)$ is the (C, α) mean of the nth partial sum of the Fourier series of e^{iwt}. Then*

$$f_n^\alpha(t) \to f(t)$$

uniformly on bounded sets.

Proof. By Lemma 4.1, $F = D^p G + \sum_{i=0}^{p-1} c_i \delta^i$ where F has support in the same interval as G. By Lemma 4.2, we have

$$\langle D^p G, \sigma_{nt}^\alpha \rangle = \langle G, D^p \sigma_{nt}^\alpha \rangle (-1)^p \to \langle G, D^p e^{iwt} \rangle (-1)^p = \langle D^p G, e^{iwt} \rangle,$$

and, similarly for $\varphi_t(w) = e^{iwt}$,

$$\langle \delta^{(i)}, \sigma_{nt}^{\alpha} \rangle = \sigma_{nt}^{\alpha(i)}(0)(-1)^i \to \varphi_t^{(i)}(0)(-1)^i = \langle \delta^{(i)}, \varphi_t \rangle.$$

COROLLARY 4.4. *Let $f(t)$ be a bandlimited function of polynomial growth bandlimited to $[-\sigma', \sigma']$. Then for each $\sigma > \sigma'$*

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{\sin \sigma(t - nT)}{\sigma(t - nT)}$$

where convergence is in the sense of (C, α)-summability for some $\alpha > p$.

This is a restatement of the theorem with a change of scale from $[-\pi, \pi]$ to $[-\sigma, \sigma]$.

REFERENCES

[1] L. L. CAMPBELL, *Sampling theorem for the Fourier transform of a distribution with bounded support*, SIAM J. Appl. Math., 16 (1968), pp. 626-636.

[2] A. J. JERRI, *The Shannon Sampling Theorem—its various extension and applications: A tutorial review*, Proc. IEEE, 65 (1977), pp. 1565-1596.

[3] R. F. HOSKINS AND J. DE SOUSA PINTO, *Sampling expansions for functions band-limited in the distributional sense*, SIAM J. Appl. Math., 44 (1984), pp. 605-610.

[4] A. J. LEE, *Characterization of band-limited functions and processes*, Inform. and Control, 31 (1976), pp. 258-271.

[5] E. PFAFFELHUBER, *Sampling series for band-limited generalized functions*, IEEE Trans. Inform. Theory, IT-17 (1971), 650-654.

[6] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1967.

[7] C. E. SHANNON, *Communications in the presence of noise*, Proc. IRE, 37 (1949), pp. 10-21.

[8] E. T. WHITTAKER, *On the functions which are represented by the expansions of interpolation theory*, Proc. Roy. Soc. Edinburgh, 35 (1915), pp. 181-194.

[9] A. H. ZEMANIAN, *Distribution Theory and Transform Analysis*, McGraw-Hill, New York, 1965.

[10a] A. ZYGMUND, *Trigonometric Series*, Vol. I, Cambridge Press, Cambridge, 1959.

[10b] ———, *Trigonometric Series*, Vol. II, Cambridge Press, Cambridge, 1959.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

ASYMPTOTICS OF COMBINATORIAL SUMS AND THE CENTRAL LIMIT THEOREM*

PAULA S. COHEN† AND AMITAI REGEV‡

Abstract. The aim of this paper is to generalize the results of Regev [*Adv. in Math.*, 41 (1981), pp. 115–136] and to simplify the proofs by deducing them from the Central Limit Theorem of probability theory. The generalized results also yield a method for calculating certain multi-integrals, some of which seem highly nontrivial.

Key words. sums, asymptotics, multi-integrals, S_n characters

AMS(MOS) subject classifications. 60F05, 20C30

1. Introduction. Let $k>0$ be an integer and $\beta>0$ a real number. For integers $n\to\infty$, the asymptotic behavior of the sum

$$S_k^{(\beta)}(n)=\sum_{\lambda\in\Lambda_k(n)} d_\lambda^\beta,$$

where $\Lambda_k(n)=\{\lambda=(\lambda_1,\cdots,\lambda_k)\in\mathbb{Z}^k \mid \lambda_1\geqq\cdots\geqq\lambda_k\geqq 0,\ \lambda_1+\cdots+\lambda_k=n\}$ and d_λ is the number of standard Young tableaux of shape λ, was studied in [12].

This was essentially done as follows. We began with a family $\{\Lambda_k(n,\rho)\}_{\rho\in\mathbb{R}}$ (see § 5) of subsets of $\Lambda_k(n)$, chosen so that for each fixed ρ, and each λ in $\Lambda_k(n,\rho)$, the d_λ, and hence the d_λ^β, could be evaluated asymptotically as $n\to\infty$. The sums of d_λ^β over λ in $\Lambda_k(n,\rho)$, converging to $S_k^{(\beta)}(n)$ as $\rho\to\infty$, were then approximated by certain integrals from which we obtained the asymptotic values of the $S_k^{(\beta)}(n)$.

Examples of such sums, having their origin in combinatorics and in algebras that satisfy polynomial identities, are given in § 5 (see also [2]).

In this paper the set $\Lambda_k(n)$ is replaced by

$$A_k(n)=\{\alpha=(\alpha_1,\cdots,\alpha_k)\in\mathbb{Z}^k \mid \alpha_1\geqq 0,\cdots,\alpha_k\geqq 0;\ \alpha_1+\cdots+\alpha_k=n\}.$$

We denote

$$\binom{n}{\boldsymbol{\alpha}}=\frac{n!}{\prod_{j=1}^k \alpha_j!}$$

and we study the asymptotics of sums of the form

$$\sum_{\boldsymbol{\alpha}\in A_k(n)} f(\boldsymbol{\alpha})\binom{n}{\boldsymbol{\alpha}}^\beta$$

for certain functions $f:\bigcup_{n\geqq 0} A_k(n)\to\mathbb{R}$.

In § 5 we show that these latter sums generalize the sums $S_k^{(\beta)}(n)$.

In § 2 we state the main result for these generalized sums, the proof of which is given in §§ 3 and 4. In particular, in § 3 we reduce the proof for arbitrary $\beta>0$ to the case $\beta=1$, and this case we treat in § 4 using the Central Limit Theorem of probability theory.

* Received by the editors November 17, 1985; accepted for publication (in revised form) August 21, 1987.

† Institut des Hautes Etudes Scientifiques, 35, Route de Chartes, Bures-sur-Yvette, 91440, France, and the Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel. This author was the recipient of a Sir Charles Clore Post-Doctoral Fellowship and Titulaire of the Centre National de la Recherche Scientifique, U.A. 763/747, Paris.

‡ Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel, and the Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802. The work of this author was supported in part by a National Science Foundation grant.

The approach of this article is much simpler than that of [12] and it allows us to deduce a general theorem about the asymptotics of such sums. All asymptotics of related sums known to us are particular cases of that theorem. Our general results also yield a method for calculating certain multi-integrals, some of which seem highly nontrivial. In § 5 we give illustrative examples of such applications.

Recently, Macdonald found some very interesting identities that involve certain multi-integrals [7]. The evaluation of these integrals is done by the Selberg formula [15]. Some of the integrals evaluated in § 5 constitute partial generalizations of these "Macdonald" (or "Mehta") integrals but it seems that their evaluation cannot be obtained from the Selberg integral.

2. The main result. For positive integers k, n we write

$$A_k(n) = \{\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k) \in \mathbb{Z}^k \,|\, \alpha_1 \geqq 0, \cdots, \alpha_k \geqq 0;\ \alpha_1 + \cdots + \alpha_k = n\},$$

$$A_k = \bigcup_{n \geqq 0} A_k(n).$$

For $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k)$ in $A_k(n)$, let

$$c_j = c_j(\boldsymbol{\alpha}) = \frac{1}{\sqrt{n}}\left(\alpha_j - \frac{n}{k}\right), \qquad j = 1, \cdots, k,$$

and $c(\boldsymbol{\alpha}) = (c_1, \cdots, c_k)$.

Given a real number $\rho > 0$ we write

$$C_k(\rho) = \{(c_1, \cdots, c_k) \in \mathbb{R}^k \,|\, |c_j| \leqq \rho, j = 1, \cdots, k\}$$

and

$$A_k(n, \rho) = \{\boldsymbol{\alpha} \in A_k(n) \,|\, c(\boldsymbol{\alpha}) \in C_k(\rho)\}.$$

DEFINITION. A function $h : A_k \to \mathbb{R}$ is defined to be permissible if

(i) There exists a polynomial $p(\mathbf{x}) = p(x_1, \cdots, x_k)$ such that for all n and all $\boldsymbol{\alpha}$ in $A_k(n)$

$$|h(\boldsymbol{\alpha})| \leqq |p(c(\boldsymbol{\alpha}))|.$$

(ii) Given $\rho > 0$,

$$\lim_{\substack{n \to \infty \\ \boldsymbol{\alpha} \in A_k(n,\rho)}} h(\boldsymbol{\alpha}) = 1.$$

By this we mean the following: given $\varepsilon > 0$, there exists an integer $N = N(\varepsilon, \rho)$ such that if $n \geqq N$ then

$$|h(\boldsymbol{\alpha}) - 1| < \varepsilon$$

for all $\boldsymbol{\alpha} \in A_k(n, p)$.

Throughout this article, for functions $f : A_k \to \mathbb{R}$, $g : \mathbb{R}^k \to \mathbb{R}$, and γ a real number the notation

$$f(\boldsymbol{\alpha}) \approx g(c(\boldsymbol{\alpha})) \cdot n^{\gamma}, \qquad \boldsymbol{\alpha} \text{ in } A_k(n),$$

means that there exists a permissible function $h : A_k \to \mathbb{R}$ and real numbers $\theta > 0$, $N_0 > 0$ such that for $n \geqq N_0$ and all $\boldsymbol{\alpha}$ in $A_k(n)$

$$|f(\boldsymbol{\alpha}) - h(\boldsymbol{\alpha}) g(c(\boldsymbol{\alpha})) \cdot n^{\gamma}| < n^{\gamma - \theta}.$$

For technical reasons we introduced the above relation $\approx$ rather than the conventional $\simeq$: recall $a_n \simeq b_n$ if $\lim_{n\to\infty}(a_n/b_n)=1$. The results of subsequent sections include the fact that if $f(\boldsymbol{\alpha}) \approx g(c(\boldsymbol{\alpha}))\cdot n^{\gamma}$, then

$$\sum_{\boldsymbol{\alpha}\in A_k(n)} f(\boldsymbol{\alpha})\binom{n}{\boldsymbol{\alpha}}^{\beta} \underset{n\to\infty}{\simeq} \sum_{\boldsymbol{\alpha}\in A_k(n)} g(c(\boldsymbol{\alpha}))\cdot n^{\gamma}\cdot\binom{n}{\boldsymbol{\alpha}}^{\beta}.$$

Remark. Let $p(x_1,\cdots,x_k)$ be a polynomial. By choosing ρ large enough, the integral

$$\int\cdots\int_{\mathbb{R}^k\setminus C_k(\rho)} |p(x_1,\cdots,x_k)|\cdot\exp\left(-\sum x_i^2\right) d^{(k)}x$$

can be made arbitrarily small. This is the only fact about $p(x)$ which will be used in the sequel (see § 4).

For this reason, one could weaken the condition that $p(x)$ is a polynomial, by requiring that $p(x)$ be a function which satisfies the above property.

The main result of the present article is as follows.

THEOREM 1. *Let γ be a real number and*

$$f: A_k \to \mathbb{R}, \qquad g: \mathbb{R}^k \to \mathbb{R}$$

be functions such that g is continuous almost everywhere and

$$f(\boldsymbol{\alpha}) \approx g(c(\boldsymbol{\alpha}))\cdot n^{\gamma}, \qquad \boldsymbol{\alpha} \text{ in } A_k(n).$$

Then for $\beta>0$ real

$$\lim_{n\to\infty}\sum_{\boldsymbol{\alpha}\in A_k(n)} f(\boldsymbol{\alpha})\binom{n}{\boldsymbol{\alpha}}^{\beta} n^{-\gamma+(1/2)(\beta-1)(k-1)}k^{-\beta n}$$
$$=\left(\frac{1}{2\pi}\right)^{(1/2)\beta(k-1)} k^{(1/2)\beta k}\int_{x_1+\cdots+x_k=0}\int_{\mathbb{R}^{k-1}} g(\mathbf{x})$$
$$\times\exp\left(-\tfrac{1}{2}\beta k(x_1^2+\cdots+x_k^2)\right) dx_1\cdots dx_{k-1}$$

whenever the integral on the right exists.

In § 3 we reduce the proof of Theorem 1 to the case $\beta=1$, which we in turn handle in § 4 using the Central Limit Theorem of probability theory.

We will also encounter the following easy variants of Theorem 1.

Variation 1. In the situation of Theorem 1, let D be a fixed domain in $\mathbb{R}^k$ and let

$$D(n)=\{\boldsymbol{\alpha}\in A_k(n)\,|\,c(\boldsymbol{\alpha})\in D\}.$$

A simple modification of the proof of Theorem 1 yields

$$\lim_{n\to\infty}\sum_{\boldsymbol{\alpha}\in A_k(n)\cap D(n)} f(\boldsymbol{\alpha})\binom{n}{\boldsymbol{\alpha}}^{\beta} n^{-\gamma+(1/2)(\beta-1)(k-1)}k^{-\beta n}$$
$$=\left(\frac{1}{2\pi}\right)^{(1/2)\beta(k-1)} k^{(1/2)\beta k}\int_{\substack{x_1+\cdots+x_k=0\\(x_1,\cdots,x_k)\in D}}\int_{\mathbb{R}^{k-1}} g(\mathbf{x})$$
$$\times\exp\left(-\tfrac{1}{2}\beta k(x_1^2+\cdots+x_k^2)\right) dx_1\cdots dx_{k-1}$$

whenever the integral on the right exists.

Variation 2. In the situation of Theorem 1 we assume further that the $g = g(x_1, \cdots, x_k)$ is a function of the differences $x_i - x_j$, $i, j = 1, \cdots, k$. We may argue as in Lemma 4.3 of [12] to obtain

$$\int_{x_1+\cdots+x_k=0} \int_{\mathbb{R}^{k-1}} g(\mathbf{x}) \exp\left(-\tfrac{1}{2}\beta k(x_1^2+\cdots+x_k^2)\right) dx_1 \cdots dx_{k-1}$$

$$= \sqrt{\frac{\beta}{2\pi}} \int \cdots \int_{\mathbb{R}^k} g(\mathbf{x}) \exp\left(-\tfrac{1}{2}\beta k(x_1^2+\cdots+x_k^2)\right) dx_1 \cdots dx_k.$$

Given the result of Theorem 1 we thereby have

$$\lim_{n\to\infty} \sum_{\boldsymbol{\alpha}\in A_k(n)} f(\boldsymbol{\alpha}) \binom{n}{\boldsymbol{\alpha}}^{\beta} n^{-\gamma+(1/2)(\beta-1)(k-1)} k^{-\beta n}$$

$$= \left(\frac{1}{2\pi}\right)^{(1/2)\beta(k-1)} k^{(1/2)\beta k} \sqrt{\frac{\beta}{2\pi}} \int \cdots \int_{\mathbb{R}^k} g(\mathbf{x})$$

$$\times \exp\left(-\tfrac{1}{2}\beta k(x_1^2+\cdots+x_k^2)\right) dx_1 \cdots dx_k$$

whenever the integral on the right exists.

We conclude this section with some illustrative examples of the types of functions we encounter in the proof of Theorem 1 and its applications.

We remark that a product of permissible functions is itself a permissible function.

Example 1. For $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k)$ in $A_k(n)$ write $\alpha_j = (n/k)(1 + c_j k/\sqrt{n})$, $j = 1, \cdots, k$, as above and suppose $\alpha_j \geqq 1$, $j = 1, \cdots, k$. If n is large enough it is easy to verify that

$$(2 + c_j^2 k)^{-1} \leqq (1 + c_j k/\sqrt{n}) \leqq k.$$

It follows that the functions $h^{(j)}: A_k \to \mathbb{R}$ and $\bar{h}^{(j)}: A_k \to \mathbb{R}$ given by

$$h^{(j)}(\boldsymbol{\alpha}) = 1 + c_j k/\sqrt{n}, \qquad \boldsymbol{\alpha} \text{ in } A_k(n), \quad j = 1, \cdots, k,$$

$$\bar{h}^{(j)}(\boldsymbol{\alpha}) = (h^{(j)}(\boldsymbol{\alpha}))^{-1}, \qquad \boldsymbol{\alpha} \text{ in } A_k(n), \quad j = 1, \cdots, k,$$

are permissible.

Example 2. For b a real number, one form of Stirling's formula is given by

$$\Gamma(n + b + 1) \simeq \sqrt{2\pi}\, e^{-n} n^{n+b} \sqrt{n}, \qquad n \to \infty.$$

Write

$$\Gamma(n + b + 1) = h_b(n) \sqrt{2\pi}\, e^{-n} n^{n+b} \sqrt{n}$$

and

$$h_{j,b}^{\pm 1}(\boldsymbol{\alpha}) = h_b^{\pm 1}(\alpha_j), \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_j) \quad \text{in } A_k, \quad j = 1, \cdots, k.$$

Clearly the $h_{j,b}^{\pm 1}$, $j = 1, \cdots, k$, are permissible functions.

Example 3. For $\boldsymbol{\alpha} \in A_k(n)$ and $c_j(\boldsymbol{\alpha})$ as before, $\alpha_i - \alpha_j = [c_i(\boldsymbol{\alpha}) - c_j(\boldsymbol{\alpha})]\sqrt{n}$, and $\alpha_i - \alpha_j + j - i \approx [c_i(\boldsymbol{\alpha}) - c_j(\boldsymbol{\alpha})]\sqrt{n}$: here $h(\boldsymbol{\alpha}) \equiv 1$, and $|[\alpha_i - \alpha_j + j - i] - [c_i(\boldsymbol{\alpha}) - c_j(\boldsymbol{\alpha})]\sqrt{n}| = |j - i| < n^{(1/2)-\theta}$ for any $0 < \theta < \frac{1}{2}$ and n large enough.

Moreover, let $r = \frac{1}{2}k(k-1)$ and choose an ordering $\{x_i - x_j \,|\, 1 \leqq i < j \leqq k\} = \{y_1, \cdots, y_r\}$. Let $M(x_i - x_j \,|\, i < j) = M(y_1, \cdots, y_r)$ be any monomial in the $x_i - x_j$'s of degree d. Clearly, $M(\alpha_i - \alpha_j \,|\, i < j) = [M(c_i(\boldsymbol{\alpha}) - c_j(\boldsymbol{\alpha}) \,|\, i < j)] \cdot n^{d/2}$, and it is easy to check that $M(\alpha_i - \alpha_j + j - i \,|\, i < j) \approx [M(c_i(\boldsymbol{\alpha}) - c_j(\boldsymbol{\alpha}) \,|\, i < j)] n^{d/2}$.

For example, $D_k(x_1, \cdots, x_k) = \prod_{i<j}(x_i - x_j)$ is of degree r; hence

$$D_k(\alpha_1 + k - 1, \alpha_2 + k - 2, \cdots, \alpha_k) \approx D_k(c(\boldsymbol{\alpha})) \cdot n^{r/2} = D_k(\alpha_1, \cdots, \alpha_k) \text{ (see § 5).}$$

3. Reduction to the case $\beta = 1$. In this section we show that the proof of Theorem 1 may be reduced to the case $\beta = 1$. By Carlson's theorem (see for example [3, p. 153, § 9.2]) it suffices to prove Theorem 1 for β a positive integer. To pass from this case to the case $\beta = 1$ we show, with notation as in § 2, the following proposition.

PROPOSITION. *Let β be a positive integer. For γ a real number and*

$$f: A_k \to \mathbb{R}, \qquad g: \mathbb{R}^k \to \mathbb{R},$$

functions such that g is continuous almost everywhere and

$$f(\boldsymbol{\alpha}) \approx g(c(\boldsymbol{\alpha})) \cdot n^{\gamma}, \qquad \boldsymbol{\alpha} \text{ in } A_k(n)$$

we have

$$\sum_{\boldsymbol{\alpha} \in A_k(n)} f(\boldsymbol{\alpha}) \binom{n}{\boldsymbol{\alpha}}^{\beta} n^{-\gamma + (1/2)(k-1)(\beta-1)} k^{-\beta n} \approx \delta^{(\beta)}(k) \sum_{\mathbf{a} \in A_k(\beta n)} f^*(\mathbf{a}) \binom{\beta n}{\mathbf{a}} n^{-\gamma} k^{-\beta n}$$

where

$$\delta^{(\beta)}(k) = \left(\frac{1}{\sqrt{2\pi}}\right)^{(k-1)(\beta-1)} \beta^{(k-1)/2} k^{(1/2)k(\beta-1)}$$

and

$$f^*: A_k \to \mathbb{R} \quad \textit{satisfies}$$

$$f^*(\mathbf{a}) \approx g\left(\frac{1}{\sqrt{\beta}} c(\mathbf{a})\right) \cdot n^{\gamma}, \qquad \mathbf{a} \text{ in } A_k(\beta n).$$

We divide the proof of the proposition into two lemmas.

For β a positive integer and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_k)$ in $A_k(n)$ we denote by $\beta\boldsymbol{\alpha}$ the element $(\beta\alpha_1, \cdots, \beta\alpha_k)$ of $A_k(\beta n)$.

LEMMA 1. *Let β be a positive integer. For $\boldsymbol{\alpha} = (a_1, \cdots, \alpha_k)$ in A_k we have as $n \to \infty$,*

$$\left(\frac{n!}{\alpha_1! \cdots \alpha_k!}\right)^{\beta} \approx h_0^{(\beta)}(\boldsymbol{\alpha}) \delta_0^{(\beta)}(k) n^{-(1/2)(k-1)(\beta-1)} \left(\frac{(\beta n)!}{(\beta\alpha_1)! \cdots (\beta\alpha_k)!}\right)$$

where

$$\delta_0^{(\beta)}(k) = \left(\frac{1}{\sqrt{2\pi}}\right)^{(k-1)(\beta-1)} \beta^{(k-1)/2} k^{(1/2)k(\beta-1)}$$

and

$$h_0^{(\beta)}: A_k \to \mathbb{R}$$

is a permissible function.

Proof. In particular, for $\boldsymbol{\alpha} \in A_k(n)$, since $0! = 1! = 1$, we may assume that $\alpha_j \geqq 1$. Applying Stirling's formula as in Example 2 of § 2, respectively, to $(n!/\alpha_1! \cdots \alpha_k!)$ and $((n)!/(\beta\alpha_1)! \cdots (\beta\alpha_k)!)$ we deduce that as $n \to \infty$

$$\left(\frac{n!}{\alpha_1! \cdots \alpha_k!}\right) \approx h_1^{(\beta)}(\boldsymbol{\alpha}) \delta_1^{(\beta)}(k) n^{(\beta-1)/2} \left(\prod_{j=1}^{k} \alpha_j\right)^{-(\beta-1)/2} \left(\frac{(\beta n)!}{(\beta\alpha_1)! \cdots (\beta\alpha_k)!}\right)$$

where

$$\delta_1^{(\beta)}(k)=\left(\frac{1}{\sqrt{2\pi}}\right)^{(k-1)(\beta-1)}\beta^{(k-1)/2}$$

and $h_1^{(\beta)}$ is a permissible function.

To conclude the proof of the lemma, we apply Example 1 and the remark preceding it in § 2

$$\left(\prod_{j=1}^{k}\alpha_j\right)^{-(\beta-1)/2}\approx\left(\frac{n}{k}\right)^{-(k/2)(\beta-1)}.$$

We now note that to each $\mathbf{a}=(a_1,\cdots,a_k)$ in $A_k(\beta n)$ we may associate a unique $\boldsymbol{\alpha}=\boldsymbol{\alpha}(\mathbf{a})=(\alpha_1,\cdots,\alpha_k)$ in $A_k(n)$ satisfying

$$\beta\alpha_j\leqq a_j<\beta\alpha_j+\beta,\qquad j=1,\cdots,k-1,$$

$$\alpha_k=n-(\alpha_1+\cdots+\alpha_{k-1}).$$

Each $\boldsymbol{\alpha}$ in $A_k(n)$ is associated in this way to exactly β^{k-1} elements of $A_k(\beta n)$.

We can easily check that as $n\to\infty$

$$c_j(\mathbf{a})\approx\sqrt{\beta}\,c_j(\boldsymbol{\alpha}(\mathbf{a})),\qquad j=1,\cdots,k,$$

in the notation of § 2.

With the correspondence $\boldsymbol{\alpha}=\boldsymbol{\alpha}(\mathbf{a})$ above we show the following lemma.

LEMMA 2. *For all* $\mathbf{a}$ *in* $A_k(\beta n)$

$$\binom{\beta n}{\beta\boldsymbol{\alpha}}=h_2^{(\beta)}(\mathbf{a})\binom{\beta n}{\mathbf{a}}$$

where $\boldsymbol{\alpha}=\boldsymbol{\alpha}(\mathbf{a})$ *and* $h_2^{(\beta)}$ *is a permissible function.*

Proof. The lemma follows on applying Stirling's formula as in Example 2 of § 2 to both sides of the above equation, and from Example 1 of § 2 and the remark preceding it. The main point is that the distance between $\mathbf{a}$ and $\beta\boldsymbol{\alpha}$ is uniformly bounded for all $\mathbf{a}$ in $A_k(\beta n)$.

The proposition now follows after applying Lemmas 1 and 2, being careful to note the remarks preceding Lemma 2.

If Theorem 1 holds for $\beta=1$ we may apply it to the right-hand sum of the proposition to deduce the result of Theorem 1 for β a positive integer, and hence, by the remarks at the beginning of the section, for all real $\beta>0$.

4. Proof of the theorem in the case $\beta=1$. By the results of § 3 it suffices to consider the case $\beta=1$. This case is a straightforward application of the Central Limit Theorem of probability and its proof is due to Gideon Schechtman.

Let $\mathbf{e}_i$, $i=1,\cdots,k$, be the vector in $\mathbb{R}^k$ with ith coordinate 1 and zeros elsewhere. We introduce random variables $\mathbf{X}_1,\cdots,\mathbf{X}_n$ that take values in the set $\{e_1,\cdots,e_k\}$ with probability

$$P(\mathbf{X}_i=\mathbf{e}_j)=\frac{1}{k},\qquad i=1,\cdots,n,\quad j=1,\cdots,k.$$

For $\boldsymbol{\alpha}$ in $A_k(n)$ we have

$$P(\mathbf{X}_1+\cdots+\mathbf{X}_n=\boldsymbol{\alpha})=\binom{n}{\boldsymbol{\alpha}}k^{-n}.$$

In this setting, the sum

$$S_k(n) = n^{-\gamma} \sum_{\alpha \in A_k(n)} f(\alpha) \binom{n}{\alpha} k^{-n}$$

equals $n^{-\gamma} E(f(\mathbf{X}_1 + \cdots + \mathbf{X}_n))$ where E stands for expected value.

On the other hand, consider the expected value $E(g((1/\sqrt{n})(\mathbf{Y}_1 + \cdots + \mathbf{Y}_n)))$ where the

$$\mathbf{Y}_i = \mathbf{X}_i - \left(\frac{1}{k}, \cdots, \frac{1}{k}\right), \qquad i = 1, \cdots, n$$

are independent random variables with expected values zero. We remark that if, for $i = 1, \cdots, n$, we write $\mathbf{Y}_i = \mathbf{Y} = (y_1, \cdots, y_k)$ then $y_k = -(y_1 + \cdots + y_{k-1})$. We may therefore apply the Central Limit Theorem (see, for example, [4, Problem 6, p. 241]) in $k-1$ dimensions with expectation matrix $\Gamma = E(y_i y_j)$ $i, j = 1, \cdots, k-1$ to deduce that

$$\lim_{n \to \infty} E\left(g\left(\frac{1}{\sqrt{n}}(\mathbf{Y}_1 + \cdots + \mathbf{Y}_n)\right)\right)$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{k-1} \sqrt{\det{}^{-1} \Gamma} \int \cdots \int_{\mathbb{R}^{k-1}} g(\mathbf{x}(\mathbf{u})) \exp\left(-\tfrac{1}{2}\mathbf{u}\Gamma^{-1}\mathbf{u}^t\right) d\mathbf{u}$$

where $\mathbf{u} = (u_1, \cdots, u_{k-1})$

$$\mathbf{x} = \mathbf{x}(\mathbf{u}) = (x_1, \cdots, x_k) = (u_1, \cdots, u_{k-1}, -(u_1 + \cdots + u_{k-1})).$$

Now, if $i \neq j$, $1 \leqq i, j \leqq k$,

$$(y_i, y_j) = \begin{cases} (1 - 1/k, -1/k) & \text{with probability } 1/k, \\ (-1/k, 1 - 1/k) & \text{with probability } 1/k, \\ (-1/k, -1/k) & \text{with probability } 1 - 2/k. \end{cases}$$

Hence $E(y_i, y_j) = (2/k)(1 - 1/k)(-1/k) + (1 - 2/k)(1/k)^2 = -1/k^2$. Similarly $E(y_j^2) = (k-1)/k^2$. It therefore follows that

$$\Gamma = \frac{1}{k^2} \begin{pmatrix} k-1 & & -1 \\ & \ddots & \\ -1 & & k-1 \end{pmatrix}, \quad \det \Gamma = k^{-k}, \quad \Gamma^{-1} = k \begin{pmatrix} 2 & & 1 \\ & \ddots & \\ 1 & & 2 \end{pmatrix}$$

and $\mathbf{u}\Gamma^{-1}\mathbf{u}^t = k \sum_{j=1}^{k} x_j^2$. Therefore

$$\lim_{n \to \infty} E\left(g\left(\frac{1}{\sqrt{n}}(\mathbf{Y}_1 + \cdots + \mathbf{Y}_n)\right)\right)$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{k-1} k^{k/2} \int \cdots \int_{\mathbb{R}^{k-1}} g(\mathbf{x}) \exp\left(-\frac{k}{2} \sum_{i=1}^{k} x_i^2\right) dx_1 \cdots dx_{k-1},$$
$$x_k = -(x_1 + \cdots + x_{k-1}).$$

It is now straightforward to verify that as

$$f(\alpha) \approx g(c(\alpha)) \cdot n^{\gamma}, \qquad \alpha \quad \text{in } A_k(n),$$

we have

$$\lim_{n \to \infty} S_k(n) = \lim_{n \to \infty} E\left(g\left(\frac{1}{\sqrt{n}}(\mathbf{Y}_1 + \cdots + \mathbf{Y}_n)\right)\right)$$

so that Theorem 1 now follows immediately in the case $\beta = 1$.

5. Applications. The applications we give here require the following theorem, which follows from our main theorem and is almost equivalent to it ($\Lambda_k = \bigcup_{n\geqq 0} \Lambda_k(n)$; § 1).

THEOREM 2. *Assume* $f:\Lambda_k \to \mathbb{R}$ *satisfies* $f(\lambda) \approx g(c(\lambda)) \cdot n^{\gamma}$ *and* $g(x_1, \cdots, x_k)$ *is continuous almost everywhere. Then*

$$\sum_{\lambda\in\Lambda_k} f(\lambda)\, d_\lambda^\beta \simeq \left(\frac{1}{2\pi}\right)^{(1/2)\beta(k-1)} \cdot k^{(1/2)\beta k^2} \cdot n^{\gamma-(1/2)(\beta-1)(k-1)-(\beta/4)k(k-1)}$$

$$\cdot k^{\beta n} \cdot \int\cdots\int_{\substack{x_1+\cdots+x_k=0\\ x_1\geqq\cdots\geqq x_k}} g(x)\cdot D_k^\beta(x)\cdot \exp\left(-\frac{k\beta}{2}\cdot\sum_{j=1}^{k} x_j^2\right) d^{(k-1)}x.$$

Proof. It is well known that

$$d_\lambda = \binom{n}{\lambda}\cdot \prod_{m=1}^{k-1}\prod_{j=1}^{m}\left(\frac{\lambda_m-\lambda_{m+j}+j}{\lambda_m+j}\right).$$

As in Examples 1 and 2 of § 2 we have

$$\frac{1}{\lambda_m+j}\approx\frac{k}{n} \quad\text{and}\quad \lambda_m-\lambda_{m+j}+j\approx(c_m(\lambda)-c_{m+j}(\lambda))\sqrt{n},$$

so

$$\prod_{m=1}^{k-1}\prod_{j=1}^{n}\left(\frac{\lambda_m-\lambda_{m+j}+j}{\lambda_m+j}\right)^\beta \approx D_k^\beta(c(\lambda))\cdot n^{-(\beta/4)k(k-1)}\cdot k^{(\beta/2)k(k-1)}.$$

The proof is now a straightforward use of Theorem 1.

Remark. Theorem 2 clearly indicates that the maximum of the d_λ, $\lambda\in\Lambda_k(n)$, is obtained for $\lambda\in\Lambda_k(n,\rho)\overset{\text{def}}{=}\Lambda_k(n)\cap A_k(n,\rho)$, for some $\rho>0$ (§ 2).

We do have a rigorous algorithmic proof of this fact (see [16]). The corresponding λ can then be found as in [1].

For further applications, the following corollary is very useful.

COROLLARY. *Let* $f:\Lambda_k\to\mathbb{R}$ *and assume* $f(\lambda)\approx g(c(\lambda))\cdot n^{\gamma}$, *where* $g(x_1,\cdots,x_k)$ *is continuous almost everywhere. Let* $d_n=\sum_{\lambda\in\Lambda_k(n)} f(\lambda)d_\lambda$, $n=1,2,\cdots$, *and assume*

$$d_n \underset{n\to\infty}{\simeq} c\cdot n^p\cdot k^n \quad (c \text{ a constant}).$$

Then $p=\gamma-\frac{1}{4}k(k-1)$ *and*

$$\int\cdots\int_{\substack{x_1+\cdots+x_k=0\\ x_1\geqq\cdots\geqq x_k}} g(x)\cdot D_K(x)\cdot\exp\left(-\frac{k}{2}\left(\sum_j x_j^2\right)\right)d^{(k-1)}x=\left(\frac{1}{k}\right)^{k^2/2}\cdot(2\pi)^{(k-1)/2}\cdot c.$$

Proof. The proof follows by equating the asymptotics of d_n with that of $\sum_{\lambda\in\Lambda_K(n)} f(\lambda)d_\lambda$, which is given by the above theorem. □

To apply this corollary, one usually constructs a series $\{\chi_n\}$, χ_n is an S_n character, such that for a fixed k and for all n, $\chi_n=\sum_{\lambda\in\Lambda_k(n)} f(\lambda)\chi_\lambda$.

If f and $d_n=\deg\chi_n$ satisfy the above assumptions, then γ and the integral are determined.

Example 4. The sums $S_k^{(\beta)}(n) = \sum_{\lambda \in \Lambda_k(n)} d_\lambda^\beta$. Here $f(\lambda) = 1$ (constant), $r = 0$, $g(x) = 1$ and by Theorem 2 we have

$$S_k^{(\beta)}(n) \simeq \left(\frac{1}{2\pi}\right)^{(1/2)\beta(k-1)} \cdot k^{(1/2)\beta k^2} \cdot n^{-(\beta/4)k(k-1)-(1/2)(\beta-1)(k-1)} \cdot k^{\beta n}$$

$$\cdot \int_{\substack{x_1+\cdots+x_k=0 \\ x_1 \geqq \cdots \geqq x_k}} \cdots \int \left[D_K(x) \exp\left(-\frac{k}{2}\left(\sum_j x_j^2\right)\right)\right]^\beta d^{(k-1)}x.$$

This agrees with (F.2.10) of [12].

Example 5. Evaluate

$$\int_{\substack{x_1+\cdots+x_k=0 \\ x_1 \geqq \cdots \geqq x_k}} \cdots \int (D_k(x))^2 \, e^{-(k/2)\|x\|^2} d^{(k-1)}x = J_2(k) \qquad \left(\|x\| = \sum_{j=1}^k x_j^2\right).$$

As a special case of the corollary we now determine $J_2(k)$. (This is a special case of the Mehta integrals [8],

$$\int_{\substack{x_1+\cdots+x_k=0 \\ x_1 \geqq \cdots \geqq x_k}} \cdots \int (D_k(x))^{2z} \cdot e^{-(k/2)\|x\|^2} d^{(k-1)}x = J_{2z}(k).$$

These can be evaluated by the Selberg integral [7].)

Let $s_k(\lambda)$ denote the number of the k-semistandard tableaux. It is well known that

$$s_k(\lambda) = D_k(\boldsymbol{\alpha}(\lambda))/D_k(k, k-1, \cdots, 1) = [\Gamma(1) \cdots \Gamma(k)]^{-1} \cdot D_k(\boldsymbol{\alpha}(\lambda)).$$

The identity $k^n = \sum_{\lambda \in \Lambda_k(n)} s_k(\lambda) d_\lambda$ can be deduced from either the Knuth-Robinson-Schensted correspondence, or from the S_n-character $\chi_n = \sum_{\lambda \in \Lambda_k(n)} s_k(\lambda)\chi_\lambda$. It is known that χ_n is the character of the natural (permuting coordinates) action of S_n on $V^{\otimes n}$, $\dim V = k$; thus $\deg \chi_n = k^n$. It follows from the corollary that $J_2(k) = \Gamma(1) \cdots \Gamma(k) \cdot \sqrt{2\pi}^{\,k-1} \cdot (1/k)^{k^2/2}$. Thus by Variation 2

$$\int_{\mathbb{R}^k} (D_k(x))^2 \, e^{-(k/2)\|x\|^2} d^{(k)}x = \left(\prod_{j=1}^k j!\right) \cdot \sqrt{2\pi}^{\,k} \cdot \left(\frac{1}{k}\right)^{k^2/2}.$$

Example 6. Let $\varphi_n \odot \psi_n$ denote the inner (Kronecker) product of the S_n characters φ_n, ψ_n. Note that *no* formula (or a "rule") is (yet) known for calculations (decomposing) $\varphi_n \odot \psi_n$. Define

$$\chi_n = \left(\sum_{\lambda \in \Lambda_k(n+1)} \chi_\lambda \odot \chi_\lambda\right)\downarrow_{S_n} \quad \text{(restrict to } S_n\text{)}.$$

By [11],

$$\chi_n = \sum_{\mu \in \Lambda_{k_2}(n)} f(\mu)\chi_\mu,$$

where $f: \Lambda_{k^2} \to \mathbb{R}$ is some function which is unknown, except for the case $k = 2$. It would be very interesting to find whether or not $f(\mu)$ has the following.

Property. (a) There exists $\gamma \in \mathbb{R}$ such that $f(\mu) \approx g(\mathbf{c}(\mu)) \cdot n^\gamma$.

(b) $g(x_1, \cdots, x_{k^2})$ is continuous almost everywhere.

Note. When $k = 2$, f does satisfy that property. We discuss that case later.

Assuming now that f does satisfy that property, we obtain by Theorem 2 that

$$\deg \chi_n \approx I_{k^2} \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^{k^2-1} \cdot \left(\frac{1}{k}\right)^{k^4} \cdot n^{\gamma-(1/4)k^2(k^2-1)} \cdot k^{2n},$$

where

$$I_{k^2}(g) = I_{k^2} = \int\cdots\int_{\substack{x_1+\cdots+x_{k_2}=0 \\ x_1\geqq\cdots\geqq x_{k_2}}} g(x_1,\cdots,x_{k^2}) \cdot D_{k^2}(x) \cdot \exp\left(-\frac{k^2}{2}(x_1^2+\cdots+x_{k^2}^2)\right) d^{(k^2-1)}x.$$

On the other hand, since $\deg \chi_n = \sum_{\lambda\in\Lambda_k(n+1)} d_\lambda^2$, hence

$$\deg \chi_n \simeq \tilde{J}_2(k) \cdot \left(\frac{1}{2\pi}\right)^{k-1} \cdot k^{k^2} \cdot n^{-(k^2-1)/2} \cdot k^{2(n+1)},$$

where $\tilde{J}_2(k) = (1/2\pi)^{k^2-1} \cdot J_2(k)$, $J_2(k)$ as in Example 5. Equating, we deduce the following.

General case. Let

$$\chi_n = \left(\sum_{\lambda\in\Lambda_k(n+1)} \chi_\lambda \odot \chi_\lambda\right)\downarrow_{S_n} = \sum_{\mu\in\Lambda_{k_2}(n)} f(\mu)\chi_\mu$$

and assume $f(\mu)$ satisfies the above property; then

$$I_{k^2} = J_2(k) \cdot \sqrt{2\pi}^{\,k^2-2k+1} \cdot k^{-k^2(k^2-1)+2}.$$

Also, we must have $\gamma = \gamma(k) = \frac{1}{4}(k^2-1)(k^2-2)$.

The case $k=2$. It follows from [9] and [6] that

$$f(\mu) \approx (c_1-c_2)(c_2-c_3)(c_3-c_4) \cdot \sqrt{n}^{\,3},$$

where $\mathbf{c} = \mathbf{c}(\mu)$. Thus

$$\int\cdots\int_{\substack{x_1+\cdots+x_4=0 \\ x_1\geqq\cdots\geqq x_4}} (x_1-x_2)(x_2-x_3)(x_3-x_4) \cdot \prod_{1\leqq i<j\leqq 4} (x_i-x_j) \cdot e^{-2(x_1^2+\cdots+x_4^2)}\, dx_1\, dx_2\, dx_3$$

$$= \tilde{J}_2(2) \cdot \sqrt{2\pi} \cdot \left(\frac{1}{2}\right)^{10} = \sqrt{2} \cdot \pi \cdot \left(\frac{1}{2}\right)^{13}.$$

Remarks. I_{k^2} determines the codimensions of the $k\times k$ matrices [14], [6] and $I_{2^2} = I_4$ was therefore calculated in [13, Appendix]. This was a long and complicated calculation by recursive methods, in which a computer was also used; a sketch of it occupies three pages in [13]. Later, William Beckner showed us a very elegant calculation of I_{2^2}, which could occupy about two printed pages. The above should be viewed as an "algebraic" (and almost effortless) calculation of I_{2^2}. For a higher number ($\geqq 3$), $f(\mu)$ and $g(\mathbf{c}(\mu))$ are unknown, but any conjecture about these can be tested by the above general case.

Example 7. Let $\psi_l(n) = \sum_{\lambda\in\Lambda_l(n)} Y_l(\lambda)\chi_\lambda$ as in [10].

The case $l=3$. By (4.2) of [10] $\deg \psi_3(n) \simeq (1/\sqrt{2\pi})^2 \cdot \sqrt{3}^{\,3} \cdot 1/n \cdot 3^n$, while $Y_3(\lambda) = \min\{\lambda_1-\lambda_2, \lambda_2-\lambda_3\}+1 \approx \min\{c_1(\lambda)-c_2(\lambda), c_2(\lambda)-c_3(\lambda)\} \cdot \sqrt{n}$.

Thus, by the corollary,

$$\iint\limits_{\substack{x_1+x_2+x_3=0\\ x_1\geqq x_2\geqq x_3}} \min\{x_1-x_2, x_2-x_3\}\cdot D_3(x)\cdot e^{-(3/2)(x_1^2+x_2^2+x_3^2)}\,dx_1\,dx_2$$
$$=\left(\frac{1}{\sqrt{3}}\right)^9\cdot(\sqrt{2\pi})^2\cdot(\sqrt{3})^3\cdot\left(\frac{1}{\sqrt{2\pi}}\right)^2=\frac{1}{27}.$$

The case $l=4$. Here $\deg\psi_4(n)\simeq(1/\sqrt{2p})^2\cdot 2^4\cdot(1/\sqrt{n})^3\cdot 4^n$.
Define

$$g(x_1,\cdots,x_4)=\begin{cases}\dfrac{1}{2}(x_1-x_2)(x_2-x_3)(x_3-x_4)-\dfrac{1}{8}[\min(x_1-x_2, x_3-x_4)]\\ \qquad\cdot(x_1-x_2-x_3+x_4)^3 \quad \text{if } x_2-x_3>\left|\dfrac{x_1-x_2-x_3+x_4}{2}\right|,\\ \dfrac{1}{2}(x_2-x_3)[\min(x_1-x_2, x_3-x_4)]\\ \qquad\cdot(x_2-x_3+[\min(x_1-x_2, x_3+x_4)]) \quad \text{if } x_2-x_3\leqq\left|\dfrac{x_1-x_2-x_3+x_4}{2}\right|,\end{cases}$$

By Theorem 5 of [5], we have $Y_4(\lambda)\approx g(c_1,\cdots,c_4)\cdot\sqrt{n}^3$, where $\mathbf{c}=\mathbf{c}(\lambda)$. Thus,

$$\iiint\limits_{\substack{x_1+\cdots+x_4=0\\ x_1\geqq\cdots\geqq x_4}} g(x_1,\cdots,x_4)D_4(x)\cdot e^{-2(x_1^2+\cdots+x_4^2)}\,d^{(3)}x$$
$$=\left(\frac{1}{\sqrt{4}}\right)^{16}\cdot(\sqrt{2\pi})^3\cdot\left(\frac{1}{\sqrt{2\pi}}\right)^3\cdot 2^4=\left(\frac{1}{2}\right)^{12}=\left(\frac{1}{4}\right)^6.$$

Example 8. Young's rule and the Littlewood-Richardson rule for the outer products of S_n-characters provide many "combinatorial" indentities; these yield quite interesting integrals. We demonstrate this below.

Let $n=m^2$, $\lambda=(n+\sqrt{n}, n-\sqrt{n})\vdash(2n)$ and *define* $\chi_{3n}=\chi_\lambda\hat{\otimes}\chi_{(n)}=\chi_{(n+\sqrt{n},n-\sqrt{n})}\hat{\otimes}\chi_{(n)}$ (outer products). By Young's rule,

$$\chi_{3n}=\sum_{\substack{\mu=(\mu_1,\mu_2,\mu_3)\vdash(3n)\\ \mu_1\geqq n+\sqrt{n}\geqq\mu_2\geqq n-\sqrt{n}\geqq\mu_3}},\qquad \chi_\mu=\sum_{\mu\in\Lambda_3(3n)} f(\mu)\chi_\mu.$$

Write $\mu_j=(3n/3)+c_j\sqrt{3n}$; then $\mu_1\geqq n+\sqrt{n}\geqq\mu_2\geqq n-\sqrt{n}\geqq\mu_3$ if and only if $c_1\geqq 1/\sqrt{3}\geqq c_2\geqq -1/\sqrt{3}\geqq c_3$. Thus

$$f(\mu)\approx g(\mathbf{c}(\mu))=\begin{cases}1, & c_1\geqq 1/\sqrt{3}\geqq c_2\geqq -1/\sqrt{3}\geqq c_3,\\ 0 & \text{otherwise.}\end{cases}$$

Clearly, $g(x_1,x_2,x_3)$ is continuous almost everywhere. Thus

$$\deg\chi_{3n}\simeq\iint\limits_{\substack{x_1+x_2+x_3=0\\ x_1\geqq 1/3\geqq x_2\geqq -1/3\geqq x_3}} D_3(x)\,e^{-(3/2)(x_1^2+x_2^2+x_3^2)}\,d^{(2)}x\cdot\left(\frac{1}{\sqrt{2\pi}}\right)^2\cdot\sqrt{3}^{3^2}\cdot n^{-3/2}\cdot 3^n.$$

Now,

$$\deg \chi_{3n} = \binom{3n}{2n} \cdot d_\lambda$$

where

$$d_\lambda = \frac{2n!}{(n+\sqrt{n})!(n-\sqrt{n})!} \frac{2\sqrt{n}+1}{n+\sqrt{n}+1}.$$

Apply Stirling's formula; since

$$\lim_{n\to\infty} \left(1+\frac{1}{\sqrt{n}}\right)^{n+\sqrt{n}} \cdot \left(1-\frac{1}{\sqrt{n}}\right)^{n-\sqrt{n}} = e,$$

we easily find that

$$\deg \chi_{3n} \simeq \frac{1}{e} \cdot \frac{\sqrt{3}}{\pi} \cdot \frac{1}{n\sqrt{n}} \cdot 3^{3n}.$$

By the corollary,

$$\iint\limits_{\substack{x_1+x_2+x_3=0 \\ x_1 \geqq 1/\sqrt{3} \geqq x_2 \geqq -1/\sqrt{3} \geqq x_3}} D_3(x)\, e^{+(3/2)(x_1^2+x_2^2+x_3^2)}\, d^{(2)}x = \left(\frac{1}{\sqrt{3}}\right)^{3^2} \cdot \sqrt{2\pi}^2 \cdot \frac{1}{e} \cdot \frac{\sqrt{3}}{\pi} = \frac{2}{3^4 e}.$$

This example can easily be generalized to higher integrals.

Acknowledgment. We are indebted to G. Schechtman for his considerable help with the probabilistic aspects of this paper.

REFERENCES

[1] R. ASKEY AND A. REGEV, *Maximal degrees for Young diagrams in a strip*, European J. Combin., 5 (1984), pp. 189-191.

[2] W. BECKNER AND A. REGEV, *Asymptotics and algebraicity of some generating functions*, Adv. in Math., 65 (1987), pp. 1-15.

[3] R. P. BOAS, JR., *Entire Functions*, Academic Press, New York, 1954.

[4] L. BREIMAN, *Probability*, Addison-Wesley, Reading, MA, 1968.

[5] P. COHEN AND A. REGEV, *Asymptotic estimates of some S_n characters and the identities of the 2×2 matrices*, Comm. Algebra, 10 (1982), pp. 71-85.

[6] E. FORMANEK, *The polynomial identities of matrices*, in Algebraist's Homage: Papers in Ring Theory and Related Topics, AMS Contemporary Mathematics, 13, 1982, pp. 41-81.

[7] I. G. MACDONALD, *Some conjectures for root systems and finite reflection groups*, SIAM J. Math. Anal., 41 (1981), pp. 998-1007.

[8] M. L. MEHTA, *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York, 1967.

[9] C. PROCESI, *Computing with 2×2 matrices*, J. Algebra, 87 (1984), pp. 342-359.

[10] A. REGEV, *The polynomial identities of matrices in characteristic zero*, Comm. Algebra, 8 (1980), pp.1417-1467.

[11] ———, *The Kronecker product of S_n characters and an $A\otimes B$ theorem for Capelli Identities*, J. Algebra, 66 (1980), pp. 505-510.

[12] ———, *Asymptotic values for degrees associated with strips of Young diagrams*, Adv. in Math., 41 (1981), pp. 115-136.

[13] ———, *Combinatorial sums, identities and trace identities of the 2×2 matrices*, Adv. in Math., 46 (1982), pp. 230-240.

[14] ———, *Codimensions and trace-dimensions of matrices are asymptotically equal*, Israel J. Math., 47 (1984), pp. 246-250.

[15] A. SELBERG, *Bemerkninger om et Multipelt Integral*, Nordisk Mat. Tidsskr., 26 (1944), pp. 71-78.

[16] P. S. COHEN AND A. REGEV, *On Maximal Degrees for Young Diagrams*, preprint.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

018

q-SERIES AND ORTHOGONAL POLYNOMIALS ASSOCIATED WITH BARNES' FIRST LEMMA*

E.G. KALNINS† AND WILLARD MILLER, JR.‡

Abstract. We exploit symmetry (recurrence relation) techniques for the derivation of properties associated with families of basic hypergeometric functions. Similar methods have been used by Nikiforov, Suslov, and Uvarov. Here we apply these ideas to find new proofs of Barnes' First Lemma and some of its q-analogues. We show that these integrals correspond to the weight functions determining the orthogonality relations for Hahn, q-Hahn, and big q-Jacobi polynomials. As another example of our method we introduce a biorthogonal system of rational functions whose weight function corresponds to the q-analogue of Kummer's Theorem.

Key words. basic hypergeometric functions, orthogonal polynomials, Barnes' Lemma, biorthogonal functions

AMS(MOS) subject classifications. 33A65, 33A75, 39A10

1. Introduction.

In papers Agarwal et al. (1987), Kalnins and Miller (1987), Miller (1988), the authors have advocated the exploitation of symmetry (recurrence relation) techniques for the derivation of properties associated with families of basic hypergeometric functions, in analogy with the local Lie theory techniques for ordinary hypergeometric functions. In particular, we have used these ideas to give simple derivations of the orthogonality relations for the Askey-Wilson and Wilson polynomials (Askey and Wilson (1985), Wilson (1980)), and, in particular, simple evaluations of the weight function integrals that determine the normalizations of these polynomials. Similar techniques have been employed by Nikiforov, Suslov, and Uvarov (1985) and Nikiforov and Suslov (1986), but they have apparently not applied them to the computation of contour integrals and summation formulas.

In §2 we use recurrence relations obeyed by a family of q-Hahn polynomials to derive the complex orthogonality of these polynomials and several q-analogues of Barnes' First Lemma, including those of Watson (1910) and Askey and Roy (1987, eq. 2.8). These integrals correspond to the square of the norm of the constant polynomial 1. Expanding one of these contour integrals by residues we obtain the real orthogonality relations for the big q-Jacobi polynomials.

In §3 we carry out the corresponding computations for the limiting case $q \to 1-$ and obtain the classical Barnes' Lemma, which we now see is associated with the orthogonality of the Hahn polynomials.

In §4 we work out a simple but nontrivial example of the use of these ideas to derive biorthogonality relations for rational basic hypergeometric functions. The associated summation formula is the q-analogue of Kummer's Theorem, originally due to Andrews (1973).

*Received by the editors June 17, 1987; accepted for publication August 13, 1987.
†Mathematics Department, University of Waikato, Hamilton, New Zealand.
‡School of Mathematics and Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota 55455. The work of this author was supported in part by the National Science Foundation under grant DMS 86-00372.

This link between recurrence relations obeyed by families of special functions, orthogonality relations for the functions, associated contour integrals and summation formulas appears capable of extensive generalization.

Most of the computations in the following sections were checked with SMP.

2. q-analogues of Barnes' Lemma. We are concerned with the q-Hahn polynomials

$$\Phi_n^{a,b,c,d}(z) = {}_3\varphi_2\left(\begin{matrix} q^{-n}, & q^{n-1}abcd, & az \\ ac, & ad & \end{matrix}; q\right), \tag{2.1}$$

$$n = 0, 1, 2, \ldots$$

where the basic hypergeometric functions ${}_{p+1}\varphi_p$ are defined as usual by

$${}_{p+1}\varphi_p\left(\begin{matrix} a_1, \ldots & a_{p+1} \\ b_1, \ldots & b_p \end{matrix}; x\right) = \sum_{m=0}^{\infty} \frac{(a_1;q)_m \ldots (a_{p+1};q)_m x^m}{(b_1;q)_m \ldots (b_p;q)_m (q;q)_m}$$

and

$$(a;q)_0 = 1,$$
$$(a;q)_m = (1-a)(1-aq)\ldots(1-aq^{m-1}), \quad m \geq 1.$$

Initially we require $0 < |q| < 1$, $acd \neq 0$, and $|a|, |b|, |c|, |d| < 1$, but some of these conditions can be relaxed later. The polynomials obey the fundamental recurrence relations

$$\mu^{(a,b,c,d)}\Phi_n^{(a,b,c,d)} = \frac{q^{\frac{1}{2}}}{d}\left(1 - \frac{ad}{q}\right)\Phi_n^{(aq^{-\frac{1}{2}}, bq^{\frac{1}{2}}, cq^{\frac{1}{2}}, dq^{-\frac{1}{2}})}, \tag{2.2A}$$

$$\tau^{(a,b,c,d)}\Phi_n^{(a,b,c,d)} = \frac{a(1-q^{-n})(1-q^{n-1}abcd)}{q^{-\frac{1}{2}}(1-ad)(1-ac)}\Phi_{n-1}^{(aq^{\frac{1}{2}}, bq^{\frac{1}{2}}, cq^{\frac{1}{2}}, dq^{\frac{1}{2}})} \tag{2.2B}$$

where

$$\mu^{(a,b,c,d)} = \frac{1}{z}\left[(1 - azq^{-\frac{1}{2}})E_z^{\frac{1}{2}} - \left(1 - \frac{zq^{\frac{1}{2}}}{d}\right)E_z^{-\frac{1}{2}}\right],$$
$$\tau^{(a,b,c,d)} = \frac{1}{z}[E_z^{\frac{1}{2}} - E_z^{-\frac{1}{2}}]$$

and $E_z^{\alpha} f(z) = f(q^{\alpha} z)$. These relations follow from

$$\mu(za;q)_n = \frac{q^{\frac{1}{2}}}{d}(1 - adq^{n-1})(zaq^{\frac{1}{2}};q)_n,$$
$$\tau(za;q)_n = \frac{a}{q^{\frac{1}{2}}}(1 - q^n)(zaq^{-\frac{1}{2}};q)_{n-1}.$$

The existence of μ suggests the existence of a recurrence taking $\Phi_n^{(aq^{-\frac{1}{2}}, bq^{\frac{1}{2}}, cq^{\frac{1}{2}}, dq^{-\frac{1}{2}})}$ to $\Phi_n^{(a,b,c,d)}$. Indeed, the appropriate operator is

$$\mu^* = -\frac{q^{\frac{1}{2}}}{\rho}\mu^{(bq^{\frac{1}{2}}, aq^{-\frac{1}{2}}, dq^{-\frac{1}{2}}, cq^{\frac{1}{2}})},$$

$$\mu^*\Phi_n^{(aq^{-\frac{1}{2}}, bq^{\frac{1}{2}}, cq^{\frac{1}{2}}, dq^{-\frac{1}{2}})} \tag{2.2C}$$

$$= \frac{q^{\frac{1}{2}}}{\rho}\frac{(\frac{ad}{q} - q^{-n})(1 - bcq^n)}{c(1 - \frac{ad}{q})}\Phi_n^{(a,b,c,d)}, \quad \rho \neq 0,$$

which follows from

$$\frac{\rho}{q^{\frac{1}{2}}}\mu^*(azq^{-\frac{1}{2}};q)_n = (az;q)_n(bcq^n-1)\frac{q^{-n}}{c} + (az;q)_{n-1}(acq^{n-1}-1)(q^n-1)\frac{q^{-n}}{c}.$$

Let $w_{a,b,c,d}(z)$ be a (complex-valued) weight function and $S_{a,b,c,d}$ be the indefinite inner product space of polynomials $f(z)$ with respect to the inner product

$$(f_1,f_2)_{a,b,c,d} = \frac{1}{2\pi i}\oint_C f_1(z)f_2(z)w_{a,b,c,d}(z)\frac{dz}{z} \tag{2.3}$$

where C is a deformation of the unit circle $|z|=1$. Now consider $\mu = \mu^{(a,b,c,d)}$ and $\mu^* = -\frac{q^{\frac{1}{2}}}{\rho}\mu^{(bq^{\frac{1}{2}},aq^{-\frac{1}{2}},dq^{-\frac{1}{2}},cq^{-\frac{1}{2}})}$ as mappings

$$\begin{aligned}&\mu : S_{a,b,c,d} \to S_{aq^{-\frac{1}{2}},bq^{\frac{1}{2}},cq^{\frac{1}{2}},dq^{-\frac{1}{2}}},\\ &\mu^* : S_{aq^{-\frac{1}{2}},bq^{\frac{1}{2}},cq^{\frac{1}{2}},dq^{-\frac{1}{2}}} \to S_{a,b,c,d}\end{aligned} \tag{2.4}$$

and determine $w_{a,b,c,d}$ so that

$$(\mu f,g)_{aq^{-\frac{1}{2}},bq^{\frac{1}{2}},cq^{\frac{1}{2}},dq^{-\frac{1}{2}}} = (f,\mu^* g)_{a,b,c,d} \tag{2.5}$$

for all $f \in S_{a,b,c,d}$, $q \in S_{aq^{-\frac{1}{2}},bq^{\frac{1}{2}},cq^{\frac{1}{2}},dq^{-\frac{1}{2}}}$. Condition (2.5) yields a q-difference equation for $w_{a,b,c,d}$ with solution

$$w_{a,b,c,d}(z) = \frac{(\frac{\rho z}{d};q)_\infty(\frac{qd}{z\rho};q)_\infty(\frac{\rho c}{z};q)_\infty(\frac{qz}{c\rho};q)_\infty}{(az;q)_\infty(bz;q)_\infty(\frac{c}{z};q)_\infty(\frac{d}{z};q)_\infty} \tag{2.6}$$

where

$$(x;q)_\infty = \lim_{n\to\infty}(x;q)_n.$$

This result is a consequence of the invariance of the contour integral under the changes of variable $z \to q^{\pm\frac{1}{2}}z$, and the property $h(qz) = -h(z)/z$ where $h(z) = (z;q)_\infty(q/z;q)_\infty$.

It follows immediately that $\mu^*\mu$ is a self-adjoint operator

$$\mu^*\mu : S_{a,b,c,d} \to S_{a,b,c,d}$$

and from the recurrence relations (2.2) we have

$$\mu^*\mu\Phi_n^{(a,b,c,d)} = \lambda_n\Phi_n^{(a,b,c,d)}, \quad \lambda_n = \frac{q}{cd\rho}\left(\frac{ad}{q}-q^{-n}\right)(1-bcq^n). \tag{2.7}$$

Since eigenfunctions corresponding to distinct eigenvalues are orthogonal we have

$$(\Phi_n^{(a,b,c,d)},\Phi_m^{(a,b,c,d)})_{a,b,c,d} = 0 \quad \text{for } m \neq n. \tag{2.8}$$

Relation (2.5) for $f = g \equiv 1$ yields

$$\|1\|^2_{aq^{-\frac{1}{2}},bq^{\frac{1}{2}},cq^{\frac{1}{2}},dq^{-\frac{1}{2}}} = -\frac{d}{\rho c}\frac{1-bc}{(1-\frac{ad}{q})}\|1\|^2_{a,b,c,d} \tag{2.9}$$

where

$$\|1\|^2_{a,b,c,d} = (1,1)_{a,b,c,d}.$$

The symmetry of the weight function in (a,b) yields an additional relation of the form (2.9). Furthermore the obvious relation

$$(\Phi_1^{(a,b,c,d)},\Phi_0^{(a,b,c,d)})_{a,b,c,d} = 0,$$

the explicit expression (2.1), and the property $(1, p_n)_{a,b,c,d} = \|1\|^2_{aq^n,b,c,d}$ for $p_n(z) = (az;q)_n$, yield the relation

$$\|1\|^2_{aq,b,c,d} = \frac{(1-ad)(1-ac)}{(1-abcd)}\|1\|^2_{a,b,c,d}. \tag{2.10}$$

Again, the symmetry in (a,b) gives an additional relation. The solution of these q-difference equations is

$$\|1\|^2_{a,b,c,d} = \frac{(abcd;q)_\infty(\frac{c\rho}{d};q)_\infty(\frac{qd}{c\rho};q)_\infty}{(ad;q)_\infty(ac;q)_\infty(ac;q)_\infty(bc;q)_\infty(bd;q)_\infty}\mathscr{K}(\rho,q) \tag{2.11}$$

where $\mathscr{K}(\rho,q)$ is to be determined. In the special case $a = \rho/d$, $b = q/c\rho$ we can compute the (trivial) integral directly: $\|1\|^2_{\rho/d,q/c\rho,c,d} = 1$. Hence $\mathscr{K}(\rho,q) = (\rho;q)_\infty(q/\rho;q)_\infty/(q;q)_\infty$ and we have

$$\begin{aligned}\|1\|^2_{a,b,c,d} &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{(\frac{\rho}{d}e^{i\theta};q)_\infty(\frac{qd}{\rho}e^{-i\theta};q)_\infty(\rho ce^{i\theta};q)_\infty}{(ae^{i\theta};q)_\infty(be^{i\theta};q)_\infty(ce^{i\theta};q)_\infty}\\ &\quad\cdot\frac{(\frac{q}{c\rho}e^{i\theta};q)_\infty}{(de^{-i\theta};q)_\infty}d\theta\\ &= \frac{(abcd;q)_\infty(\frac{c\rho}{d};q)_\infty(\frac{qd}{c\rho};q)_\infty(\rho;q)_\infty(\frac{q}{\rho};q)_\infty}{(ad;q)_\infty(ac;q)_\infty(bc;q)_\infty(bd;q)_\infty(q;q)_\infty}\end{aligned} \tag{2.12}$$

in agreement with Askey and Roy (1986).

Now we consider the recurrence

$$\tau^{(a,b,c,d)} : S_{a,b,c,d} \to S_{aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12}},$$

(2.2B) and compute the adjoint $\tau^* \equiv \tau^{*(aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12})}$ such that

$$(\tau f, g)_{aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12}} = (f, \tau^* g)_{a,b,c,d} \tag{2.13}$$

for all $f \in S_{a,b,c,d}$, $g \in S_{aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12}}$. A straightforward computation yields

$$\tau^{*(aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12})} = \frac{1}{q^{\frac12}z}\left[-(1-az)(1-bz)E_z^{\frac12} + \left(1-\frac{z}{c}\right)\left(1-\frac{z}{d}\right)E_z^{-\frac12}\right]. \tag{2.14}$$

It follows that $\tau^*\tau : S_{a,b,c,d} \to S_{a,b,c,d}$ is self-adjoint. Moreover, the action of τ^* on the polynomial basis is

$$\tau^*\Phi_{n-1}^{(aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12})} = \frac{(1-ac)(1-ad)}{q^{\frac12}acd}\Phi_n^{(a,b,c,d)}. \tag{2.15}$$

This follows from

$$\tau^*(aq^{\frac12}z;q)_k = \frac{q^{-k-\frac12}}{acd}(1-acq^k)(1-adq^k)(az;q)_k - \frac{q^{-k-\frac12}}{acd}(1-abcdq^k)(az;q)_{k+1}.$$

Thus

$$\tau^*\tau\Phi_n^{(a,b,c,d)} = \frac{1}{cd}(1-q^{-n})(1-q^{n-1}abcd)\Phi_n^{(a,b,c,d)}. \tag{2.16}$$

Setting $f = \Phi_n^{(a,b,c,d)}$, $g = \Phi_{n-1}^{(aq^{\frac12},bq^{\frac12},cq^{\frac12},dq^{\frac12})}$ in (2.13) we obtain the recurrence

$$\|\Phi_n^{(a,b,c,d)}\|^2_{a,b,c,d} = \frac{qa^2cd(1-q^{-n})(1-q^{n-1}abcd)}{(1-ad)^2(1-ac)^2}\|\Phi_{n-1}^{(aq^{\frac12},\ldots,dq^{\frac12})}\|^2_{aq^{\frac12},\ldots,dq^{\frac12}}, \tag{2.17}$$

which permits us to compute the norms $\|\Phi_n^{(a,\dots,d)}\|^2_{a,\dots,d}$ recursively from $\|1\|^2_{a,\dots,d}$. Note that the norms are all nonzero. We have shown that the q–Hahn polynomials $\{\Phi_n^{(a,b,c,d)}\}$ are uniquely characterized by their orthogonality with respect to the complex weight function $w_{a,b,c,d}$.

Since the weight function is symmetric in $\{a,b\}$ the orthogonal polynomials satisfies the transformation rule

(2.18)
$$ {}_3\varphi_2\left(\begin{matrix} q^{-n}, & q^{n-1}abcd, & bz \\ bc, & dc & \end{matrix}; q\right) = \left(\frac{b}{a}\right)^n \frac{(ac;q)_n(ad;q)_n}{(bc;q)_n(bd;q)_n} {}_3\varphi_2\left(\begin{matrix} q^{-n}, & q^{n-1}abcd, & az \\ ac, & ad & \end{matrix}; q\right). $$

Also, the action of $\tau^{*(aq^{\frac12},\dots,dq^{\frac12})}$ determines a Rodrigues formula.

The complex orthogonality (2.8) for the q-Hahn polynomials leads to real discrete orthogonality for the big q-Jacobi polynomials (Andrews and Askey (1985)):

$$ {}_3\varphi_2\left(\begin{matrix} q^{-n}, & q^{n-1+\alpha+\beta+\gamma+\delta}, & q^{x+1} \\ q^{\alpha+\gamma}, & q^{\alpha+\delta} & \end{matrix}; q\right). \tag{2.19} $$

The polynomials (2.19) are orthogonal with respect to the discrete measure with mass points and corresponding weights (Ismail and Wilson (1982)):

$$ \begin{aligned} x &= \alpha+\delta+k-1 \qquad \frac{(q^{\delta-\gamma+k+1};q)_\infty (q^{k+1};q)_\infty}{(q^{\alpha+\delta+k};q)_\infty (q^{\beta+\delta+k};q)_\infty} q^k q^{\alpha+\delta-1}(1-q), \\ x &= \alpha+\gamma+k-1 \qquad \frac{(q^{\gamma-\delta+k+1};q)_\infty (q^{k+1};q)_\infty}{(q^{\alpha+\gamma+k};q)_\infty (q^{\beta+\gamma+k};q)_\infty} q^k q^{\alpha+\gamma-1}(1-q), \end{aligned} \tag{2.20} $$

$k = 0,1,2,\dots$. To obtain this result from (2.6), (2.8) we first set $a = q^\alpha, \dots, d = q^\delta$. If $\operatorname{Re}(\alpha,\beta,\gamma,\delta) > 0$ and there are no double poles we can expand the integral $(\Phi_n,\Phi_m)_{a,b,c,d}$ by residues, using the simple poles of $w_{a,b,c,d}(z)$, (2.6), at $z = q^{\gamma+k}$, $z = q^{\delta+k}$. The result of this expansion is (2.20) with $z = q^x$. In particular, the dependence on ρ cancels out.

Our approach relating orthogonal polynomials to integrals of weight functions can be used to evaluate other important integrals. One class of such integrals can be conveniently studied through the change of variables given in the preceding paragraph:

$$ a = q^\alpha, \quad b = q^\beta, \quad c = q^\gamma, \quad d = q^\delta, \quad z = q^x. \tag{2.21} $$

This change reduces q-difference equations for the weight function to ordinary difference equations. Indeed the operators (2.2) now take the form

$$ \begin{aligned} \mu^{(\alpha,\beta,\gamma,\delta)} &= q^{-x}[(1-q^{\alpha+x-\frac12})\mathscr{E}_x^{\frac12} - (1-q^{x-\delta+\frac12})\mathscr{E}_x^{-\frac12}], \\ \tau^{(\alpha,\beta,\gamma,\delta)} &= q^{-x}[\mathscr{E}_x^{\frac12} - \mathscr{E}_x^{-\frac12}], \\ \mu* &= -q^{-x}[(1-q^{\beta+x})\mathscr{E}_x^{\frac12} - (1-q^{x-\gamma})\mathscr{E}_x^{-\frac12}] \end{aligned} \tag{2.22} $$

where $\mathscr{E}_x^s g(x) = g(x+s)$.

The orthogonal functions are now polynomials in q^x. We require that the inner product take the form

$$ (f_1,f_2)_{\alpha,\beta,\gamma,\delta} = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} f_1(q^x) f_2(q^x) w_{\alpha,\beta,\gamma,\delta}(x)\,dx \tag{2.23} $$

where the contour in the complex x-plane will run from $-i\infty$ to $+i\infty$ so that decreasing sequences of poles for w lie on the left and increasing sequences of poles lie on the right. The condition that μ^* be the adjoint of μ now becomes

$$ (\mu f, g)_{\alpha-\frac12,\beta+\frac12,\gamma+\frac12,\delta-\frac12} = (f,\mu^* g)_{\alpha,\beta,\gamma,\delta} \tag{2.24} $$

with $f \in S_{\alpha,\beta,\gamma,\delta}$, $g \in S_{\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}}$, where $S_{\alpha,\beta,\gamma,\delta}$ is the space of polynomials in q^x with inner product (2.23). The nonunique solution for w is

$$w_{\alpha,\beta,\gamma,\delta}(x) = \frac{(q^{x-\gamma+1};q)_\infty (q^{x-\delta+1};q)_\infty}{(q^{x+\alpha};q)_\infty (q^{x+\beta};q)_\infty} q^x H(\alpha,\beta,\gamma,\delta,x) \tag{2.25}$$

where H is an analytic function of its variables satisfying the periodicity properties

$$\begin{aligned} H(\alpha,\beta,\gamma,\delta,x) &= H(\alpha,\beta,\gamma,\delta,x+1), \\ H\left(\alpha-\frac{1}{2},\beta-\frac{1}{2},\gamma\pm\frac{1}{2},\delta\mp\frac{1}{2},x+\frac{1}{2}\right) &= H\left(\alpha-\frac{1}{2},\beta-\frac{1}{2},\gamma\pm\frac{1}{2},\delta\mp\frac{1}{2},x+\frac{1}{2}\right) \\ &= H\left(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2},x+\frac{1}{2}\right). \end{aligned} \tag{2.26}$$

One solution of (2.26) is

$$H(\alpha,\beta,\gamma,\delta,x) = \frac{\sin\pi(\gamma-\delta)}{\sin\pi(\gamma-x)\sin\pi(\delta-x)}$$

so that the weight function becomes

$$w_{\alpha,\beta,\gamma,\delta}(x) = \frac{(q^{x-\gamma+1};q)_\infty (q^{x-\delta+1};q)_\infty q^x \sin\pi(\gamma-\delta)}{(q^{x+\alpha};q)_\infty (q^{x+\beta};q)_\infty \sin\pi(\gamma-x)\sin\pi(\delta-x)}. \tag{2.27}$$

Then the polynomials are

(2.28)

$$\Phi_n^{(\alpha,\beta,\gamma,\delta)}(q^x) = {}_3\varphi_2\left(\begin{matrix} q^{-n}, & q^{n+\alpha+\beta+\gamma+\delta-1}, & q^{\alpha+x} \\ q^{\alpha+\gamma}, & q^{\alpha+\delta} & \end{matrix}; q\right), \qquad n = 0,1,2,\ldots$$

and the eigenvalue equation is

$$\mu^*\mu\Phi_n^{\alpha,\beta,\gamma,\delta} = q^{\frac{1}{2}-\gamma-\delta}(q^{\alpha+\delta-1}-q^{-n})(1-q^{\beta+\gamma+n})\Phi_n^{(\alpha,\beta,\gamma,\delta)}.$$

We have immediately

$$(\Phi_n^{(\alpha,\ldots,\delta)}, \Phi_m^{(\alpha,\ldots,\delta)})_{\alpha,\ldots,\delta} = 0, \qquad m \neq n$$

and

$$\|1\|^2_{\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}} = q^{-\frac{1}{2}+\delta-\gamma}\frac{(1-q^{\beta+\gamma})}{(1-q^{\alpha+\delta-1})}\|1\|^2_{\alpha,\beta,\gamma,\delta}. \tag{2.29}$$

The symmetry of w in (α,β) leads to a similar recurrence. Also, the skew-symmetry of w in γ,δ yields a new recurrence. Finally, the orthogonality

$$(\Phi_1^{(a,\ldots,\delta)}, \Phi_0^{(\alpha,\ldots,\delta)}) = 0$$

leads to the recurrences

$$\|1\|^2_{\alpha+1,\beta,\gamma,\delta} = \frac{(1-q^{\alpha+\gamma})(1-q^{\alpha+\delta})}{(1-q^{\alpha+\beta+\gamma+\delta})}\|1\|^2_{\alpha,\beta,\gamma,\delta}$$

and

$$\|1\|^2_{\alpha,\beta,\gamma+1,\delta} = -\frac{q^{\delta-\gamma}(1-q^{\alpha+\gamma})(1-q^{\beta+\gamma})}{(1-q^{\alpha+\beta+\gamma+\delta})}\|1\|^2_{\alpha,\beta,\gamma,\delta}$$

with the solution

$$\|1\|^2_{\alpha,\beta,\gamma,\delta} = \frac{(q^{\alpha+\beta+\gamma+\delta};q)_\infty (q^{\delta-\gamma};q)_\infty (q^{1+\gamma-\delta};q)_\infty q^\gamma}{(q^{\alpha+\delta};q)_\infty (q^{\alpha+\gamma};q)_\infty (q^{\beta+\gamma};q)_\infty (q^{\beta+\delta};q)_\infty} M(\alpha,\beta,\gamma,\delta) \tag{2.30}$$

where M is an analytic function of its arguments, symmetric in (α,β) and in (γ,δ), satisfying the periodicity relations

$$\begin{aligned} M(\alpha,\beta,\gamma,\delta) &= M(\alpha+1,\beta,\gamma,\delta) = M(\alpha,\beta,\gamma+1,\delta) \\ &= M\left(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}\right). \end{aligned} \tag{2.31}$$

To evaluate M we replace α by $\alpha+k$, k a positive integer, and rewrite the integral for $\|1\|^2_{\alpha+k,\beta,\gamma,\delta}$ in the form $(x=iy)$

$$\begin{aligned} &\frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\frac{(q^{\alpha+k+\delta};q)_\infty (q^{\alpha+k+\delta};q)_\infty}{(q^{\alpha+\beta+\gamma+\delta+k};q)_\infty (q^{iy+\alpha+k};q)_\infty}\right]\frac{(q^{iy-\gamma+1};q)_\infty}{(q^{iy+\beta};q)_\infty} \\ &\qquad \cdot\frac{(q^{iy-\delta+1};q)_\infty q^{iy}dy}{\sin\pi(\gamma-iy)\sin\pi(\delta-iy)} \\ &= \frac{q^\gamma (q^{\delta-\gamma};q)_\infty (q^{1+\gamma-\delta};q)_\infty}{(q^{\beta+\gamma};q)_\infty (q^{\beta+\delta};q)_\infty} M(\alpha,\beta,\gamma,\delta). \end{aligned} \tag{2.32}$$

Notice that the right-hand side of (2.32) is independent of k and that the bracketed quantity on the left goes to 1 as $k\to+\infty$. From the Lebesgue dominated convergence theorem we conclude that

$$\begin{aligned} &\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{(q^{iy-\gamma+1};q)_\infty (q^{iy-\delta+1};q)_\infty q^{iy}}{(q^{iy+\beta};q)_\infty \sin\pi(\gamma-iy)\sin\pi(\delta-iy)} \\ &= \frac{q^\delta (q^{\delta-\gamma};q)_\infty (q^{1+\gamma-\delta};q)_\infty}{(q^{\beta+\gamma};q)_\infty (q^{\beta+\delta};q)_\infty} M(\alpha,\beta,\gamma,\delta). \end{aligned}$$

It follows immediately that M is independent of α and β. Now in (2.32) set $k=0$, $\alpha=1-\gamma$, $\beta=1-\delta$:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{q^{iy}\sin\pi(\gamma-\delta)dy}{\sin\pi(\gamma-iy)\sin\pi(\delta-iy)} = \frac{M(\gamma,\delta)q^\gamma(1-q^{\delta-\gamma})}{(1-q)(q;q)_\infty}. \tag{2.33}$$

The rather elementary integral on the left-hand side of (2.33) can be easily evaluated by residues and the resulting geometric series summed to yield $M(\gamma,\delta)=(q;q)_\infty/\pi$. Thus

$$\begin{aligned} &\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{(q^{iy-\gamma+1};q)_\infty (q^{iy-\delta+1};q)_\infty q^{iy}\sin\pi(\gamma-\delta)}{(q^{iy+\alpha};q)_\infty (q^{iy+\beta};q)_\infty \sin\pi(\gamma-iy)\sin\pi(\delta-iy)}dy \\ &= \frac{(q;q)_\infty q^\gamma (q^{\alpha+\beta+\gamma+\delta};q)_\infty (q^{\delta-\gamma};q)_\infty (q^{1+\gamma-\delta};q)_\infty}{\pi(q^{\alpha+\delta};q)_\infty (q^{\alpha+\gamma};q)_\infty (q^{\beta+\gamma};q)_\infty (q^{\beta+\delta};q)_\infty}, \end{aligned} \tag{2.34}$$

which is Watson's q-analogue of Barnes' First Lemma (Watson (1910)).

By choosing other solutions H of relations (2.26) we can evaluate other integrals in the form (2.30), (2.31). For example, if $H=\cos^{-2}\pi x$ then $M=(q;q)_\infty q^{\frac{1}{4}}(1-q^{\frac{1}{2}})/2\pi(q^\gamma-q^\delta)$. However, in general one cannot evaluate M by the simple method we used for (2.34).

3. The classical Barnes' Lemma. To see the relationship between our results and Barnes' Lemma we could, with care, let $q \to 1-$ in expression (2.12) (i.e., at the end of our construction). However, it is more instructive to take the limit $q \to 1-$ immediately and then proceed step-by-step through the argument of the preceding section. From this point of view the functions to be considered are the Hahn polynomials

$$\Phi_n^{(\alpha,\beta,\gamma,\delta)}(x) = {}_3F_2\left(\begin{matrix} -n, & n+\alpha+\beta+\gamma+\delta-1, & \alpha+x \\ \alpha+\gamma, & \alpha+\delta & \end{matrix}; 1\right) \tag{3.1}$$

$$n = 0, 1, 2, \ldots, \qquad \alpha, \beta, \gamma, \delta > 0$$

where ${}_3F_2$ is a generalized hypergeometric function:

$${}_{p+1}F_p\left(\begin{matrix} \alpha, \ldots, \alpha_{p+1} \\ \beta_1, \ldots, \beta_p \end{matrix}, z\right) = \sum_{m=0}^{\infty} \frac{(\alpha_1)_m \ldots (\alpha_{p+1})_m z^m}{(\beta_1)_m \ldots (\beta_p)_m m!},$$

$$(\alpha)_m = \begin{cases} 1 & \text{if } m = 0, \\ \alpha(\alpha+1)\ldots(\alpha+m-1) & \text{if } m \geq 1. \end{cases}$$

The recurrence relations are

$$\mu^{(\alpha,\beta,\gamma,\delta)}\Phi_n^{(\alpha,\beta,\gamma,\delta)} = (\alpha+\delta-1)\Phi_n^{(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2})}, \tag{3.2A}$$

(3.2B)

$$\mu^{*(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2})}\Phi_n^{(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2})} = \frac{(n+\alpha+\delta-1)(n+\beta+\gamma)}{(\alpha+\delta-1)}\Phi_n^{(\alpha,\beta,\gamma,\delta)},$$

$$\tau^{(\alpha,\beta,\gamma,\delta)}\Phi_n^{(\alpha,\beta,\gamma,\delta)} = \frac{-n(n+\alpha+\beta+\gamma+\delta-1)}{(\alpha+\delta)(\alpha+\gamma)}\Phi_{n-1}^{(\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2})}, \tag{3.2C}$$

$$\tau^{*(\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2})}\Phi_{n-1}^{(\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2})} = (\alpha+\delta)(\alpha+\gamma)\Phi_n^{(\alpha,\beta,\gamma,\delta)}, \tag{3.2D}$$

where

$$\mu^{(\alpha,\beta,\gamma,\delta)} = \left(\alpha+x-\frac{1}{2}\right)\mathscr{E}_x^{\frac{1}{2}} + \left(\delta-x-\frac{1}{2}\right)\mathscr{E}_x^{-\frac{1}{2}},$$

$$\mu^{*(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2})} = (\beta+x)\mathscr{E}_x^{\frac{1}{2}} + (\gamma-x)\mathscr{E}_x^{-\frac{1}{2}},$$

$$\tau^{(\alpha,\beta,\gamma,\delta)} = \mathscr{E}_x^{\frac{1}{2}} - \mathscr{E}_x^{-\frac{1}{2}},$$

$$\tau^{*(\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2})} = (\gamma-x)(\delta-x)\mathscr{E}_x^{-\frac{1}{2}} - (\alpha+x)(\beta+x)\mathscr{E}_x^{\frac{1}{2}},$$

and $\mathscr{E}_x^s g(x) = g(x+s)$.

We define a complex inner product by

$$(g_1, g_2)_{\alpha,\beta,\gamma,\delta} = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} g_1(x)g_2(x)w_{\alpha,\beta,\gamma,\delta}(x)dx \tag{3.3}$$

for polynomials g_1, g_2 where the integration path is the imaginary axis in the complex x-plane. Let $S_{\alpha,\beta,\gamma,\delta}$ be the space of polynomials with this inner product. We require that

$$(\mu f, g)_{\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}} = (f, \mu^* g)_{\alpha,\beta,\gamma,\delta} \tag{3.4}$$

for all $f \in S_{\alpha,\beta,\gamma,\delta}$, $g \in S_{\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}}$. In order that (3.4) hold, the weight function must satisfy a difference equation whose solution is (essentially)

$$w_{\alpha,\beta,\gamma,\delta}(x) = \Gamma(\alpha+x)\Gamma(\beta+x)\Gamma(\gamma-x)\Gamma(\delta-x) \tag{3.5}$$

where Γ is the gamma function (Whittaker and Watson (1958, Chap. XII)). Here we are using the fundamental recurrence for the gamma function

$$\Gamma(z+1) = z\Gamma(z).$$

It now follows that $\mu^*\mu : S_{\alpha,\beta,\gamma,\delta} \to S_{\alpha,\beta,\gamma,\delta}$ is self-adjoint with respect to this inner product and has eigenfunctions $\Phi_n^{\alpha,\beta,\gamma,\delta}$:

$$\mu^*\mu\Phi_n^{\alpha,\beta,\gamma,\delta} = (n+\alpha+\delta-1)(n+\beta+\gamma)\Phi_n^{\alpha,\beta,\gamma,\delta}. \tag{3.6}$$

Since eigenfunctions corresponding to distinct eigenvalues are orthogonal we have

$$(\Phi_n^{(\alpha,\beta,\gamma,\delta)}, \Phi_m^{(\alpha,\beta,\gamma,\delta)})_{\alpha,\beta,\gamma,\delta} = 0, \qquad n \neq m. \tag{3.7}$$

A similar computation gives

$$(\tau f, g)_{\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2}} = (f, \tau^* g)_{\alpha,\beta,\gamma,\delta} \tag{3.8}$$

for the same weight function and all

$$g \in S_{\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2}}, \qquad f \in S_{\alpha,\beta,\gamma,\delta}.$$

Thus $\tau^*\tau$ is self-adjoint on $S_{\alpha,\beta,\gamma,\delta}$ and

$$\tau^*\tau\Phi_n^{(\alpha,\beta,\gamma,\delta)} = -n(n+\alpha+\beta+\gamma+\delta-1)\Phi_n^{(\alpha,\beta,\gamma,\delta)}. \tag{3.9}$$

Setting $f = g = 1$ in (3.4) we find

$$\|1\|^2_{\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}} = \frac{(\beta+\gamma)}{(\alpha+\delta-1)}\|1\|^2_{\alpha,\beta,\gamma,\delta}. \tag{3.10}$$

Symmetry of the weight function in (α,β) and in (γ,δ) gives three more such relations. Furthermore

$$(\Phi_1^{(\alpha,\beta,\gamma,\delta)}, \Phi_0^{(\alpha,\beta,\gamma,\delta)})_{\alpha,\beta,\gamma,\delta} = 0,$$

which, from (3.1) and (3.5), implies

$$\|1\|^2_{\alpha+1,\beta,\gamma,\delta} = \frac{(\alpha+\gamma)(\alpha+\delta)}{(\alpha+\beta+\gamma+\delta)}\|1\|^2_{\alpha,\beta,\gamma,\delta} \tag{3.11}$$

and also

$$\|1\|^2_{\alpha,\beta,\gamma+1,\delta} = \frac{(\gamma+\alpha)(\gamma+\beta)}{(\alpha+\beta+\gamma+\delta)}\|1\|^2_{\alpha,\beta,\gamma,\delta}. \tag{3.12}$$

The symmetry of the weight function in (α,β) and in (γ,δ) gives two more such relations. It follows that

$$\|1\|^2_{\alpha,\beta,\gamma,\delta} = \frac{\Gamma(\alpha+\gamma)\Gamma(\alpha+\delta)\Gamma(\beta+\gamma)\Gamma(\beta+\delta)}{\Gamma(\alpha+\beta+\gamma+\delta)}M(\alpha,\beta,\gamma,\delta) \tag{3.13}$$

where M is symmetric in (α,β) and in (γ,δ), and satisfies the periodicity properties

$$\begin{aligned} M(\alpha+1,\beta,\gamma,\delta) &= M\left(\alpha-\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta-\frac{1}{2}\right) \\ &= M(\alpha,\beta,\gamma+1,\delta) = M(\alpha,\beta,\gamma,\delta). \end{aligned} \tag{3.14}$$

To evaluate M we replace α by $\alpha + k$ and γ by $\gamma + k$, k a positive integer, and write the expression (3.13) in the form

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\frac{\Gamma(\alpha+\beta+\gamma+\delta+2k)\Gamma(\alpha+k+iy)\Gamma(\gamma+k-iy)}{\Gamma(\alpha+\gamma+2k)\Gamma(\alpha+\delta+k)\Gamma(\beta+\gamma+k)}\right)$$
$$\cdot\,\Gamma(\beta+iy)\Gamma(\delta-iy)dy = \Gamma(\beta+\delta)M(\alpha,\beta,\gamma,\delta). \tag{3.15}$$

From Stirling's formula (Whittaker and Watson (1958, Chap. XII)), we have

$$\lim_{k\to+\infty}\frac{\Gamma(\alpha+\beta+\gamma+\delta+2k)\Gamma(\alpha+k+iy)\Gamma(\gamma+k-iy)}{\Gamma(\alpha+\gamma+2k)\Gamma(\alpha+\delta+k)\Gamma(\beta+\gamma+k)} = 2^{\beta+\delta}$$

and it follows easily that

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\Gamma(\beta+iy)\Gamma(\delta-iy)dy = 2^{-(\beta+\delta)}\Gamma(\beta+\delta)M.$$

Hence, M is a constant, independent of $\alpha,\beta,\gamma,\delta$. To evaluate the constant we set $\alpha = \beta+\gamma = \delta = \frac{1}{2}$ in (3.13) and use the reflection formula for gamma functions:

$$\|1\|^2_{\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}} = M = 2\pi\int_0^{\infty}\frac{dy}{\cosh\pi y} = 1.$$

Thus,

$$\|1\|^2_{\alpha,\beta,\gamma,\delta} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\Gamma(\alpha+iy)\Gamma(\beta+iy)\Gamma(\gamma-iy)\Gamma(\delta-iy)dy$$
$$= \frac{\Gamma(\alpha+\gamma)\Gamma(\alpha+\delta)\Gamma(\beta+\gamma)\Gamma(\beta+\delta)}{\Gamma(\alpha+\beta+\gamma+\delta)}. \tag{3.16}$$

This is Barnes' First Lemma (Bailey (1935, p.6), Slater (1966, p. 109)).

In relation (3.8) we set $f = \Phi_n^{(\alpha,\beta,\gamma,\delta)}$, $g = \Phi_{n-1}^{(\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2})}$ to obtain the recurrence

$$\|\Phi_n^{(\alpha,\beta,\gamma,\delta)}\|^2_{\alpha,\beta,\gamma,\delta} = \frac{-n(n+\alpha+\beta+\gamma+\delta-1)}{(\alpha+\gamma)^2(\alpha+\delta)^2}\cdot$$
$$\cdot\,\|\Phi_{n-1}^{(\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2})}\|^2_{\alpha+\frac{1}{2},\beta+\frac{1}{2},\gamma+\frac{1}{2},\delta+\frac{1}{2}}. \tag{3.17}$$

It follows that the norms of the orthogonal polynomials are nonzero and can be computed recursively from $\|1\|^2_{\alpha,\beta,\gamma,\delta}$. Thus these polynomials are defined uniquely by their orthogonality with respect to the weight function w.

The symmetry of the weight function in (α,β) implies the identity

$${}_3F_2\left(\begin{matrix}-n, & n+\alpha+\beta+\gamma+\delta-1, & x+\beta\\ \beta+\gamma, & \beta+\delta & \end{matrix};1\right)$$
$$= \frac{(\alpha+\gamma)_n(\alpha+\delta)_n}{(\beta+\gamma)_n(\beta+\delta)_n}{}_3F_2\left(\begin{matrix}-n, & n+\alpha+\beta+\gamma+\delta-1, & x+\alpha\\ \alpha+\gamma, & \alpha+\delta & \end{matrix};1\right).$$

Furthermore, the symmetry of the weight function with respect to the interchanges $x\leftrightarrow -x$, $\alpha\leftrightarrow\delta$, $\beta\leftrightarrow\gamma$ implies

$${}_3F_2\left(\begin{matrix}-n, & n+\alpha+\beta+\gamma+\delta-1, & -x+\alpha\\ \alpha+\gamma, & \alpha+\delta & \end{matrix};1\right)$$
$$= (-1)^n\frac{(\beta+\delta)_n}{(\alpha+\gamma)_n}{}_3F_2\left(\begin{matrix}-n, & n+\alpha+\beta+\gamma+\delta-1, & x+\delta\\ \alpha+\delta, & \beta+\delta & \end{matrix};1\right).$$

4. Biorthogonality relations. We will extend the ideas of the previous sections by considering rational functions rather than polynomials. Thus the basic object of study will be the rational function of z

$$p_n^{a,b}(z) = \frac{(az;q)_n}{(bz;q)_n} \tag{4.1}$$

rather than the polynomial $(az;q)_n$ of §2. Two fundamental recurrences are

$$\mu_1 p_{n-1}^{a,bq} = \frac{(a - \frac{b\rho}{a}q^{-n+1})}{a-b} p_{n-2}^{a,b} + \frac{b}{a}\frac{(-b+\rho q^{-n+1})}{a-b} p_n^{a,b}, \tag{4.2A}$$

$$\mu_2 p_n^{a,b} = (a-b)(1-q^n) p_{n-1}^{a,bq}, \tag{4.2B}$$

where

$$\mu_1 = \frac{(1 - \frac{b\rho z}{aq})}{(1 - \frac{az}{q})} E_z^{-1} + \frac{b}{a}\frac{1-\rho z}{1-bz} I_z,$$

$$\mu_2 = \frac{q}{z}(1 - bq^{-1}z)(1-bz)[I_z - E_z^{-1}],$$

and

$$E_z^s f(z) = f(q^s z), \quad I_z f(z) = f(z).$$

It is not difficult to verify that the eigenvalue equation

$$\mu_1 \mu_2 \Psi(z) = \lambda \Psi(z) \tag{4.3}$$

has the solutions

$$\Psi_\ell^{a,b,\rho}(z) = \sum_{k=0}^{\ell} \frac{(q^{-2\ell};q^2)_k (\frac{b}{\rho}q^{2\ell-1};q^2)_k}{(q^2;q^2)_k (\frac{a^2 q}{b\rho};q^2)_k} \frac{(az;q)_{2k}}{(bz;q)_{2k}} q^{2k} \tag{4.4}$$

corresponding to the eigenvalues

$$\lambda_\ell = \frac{b}{a}(1-q^{-2\ell})(bq^{2\ell} - \rho q), \quad \ell = 0,1,2,\ldots. \tag{4.5}$$

Let $S_e^{a,b,\rho}$ be the complex vector space of all finite linear combinations of the functions $\{\Psi_\ell^{a,b,\rho}\}$. Consider the bilinear form

$$< f, g >_{a,b,\rho}^e = \oint_C f(z) g(z) w_{a,b,\rho}^e(z) \frac{dz}{z} \tag{4.6}$$

where $g \in S_e^{a,b,\rho}$, w is a weight function, C is a positively oriented closed curve in the complex z-plane and $f \in \hat{S}_e^{a,b,\rho}$ (a space to be determined). We interpret $\mu_1\mu_2$ as the map

$$\mu_1\mu_2 : S_e^{a,b,\rho} \to S_e^{a,b,\rho}$$

and try to determine w, C, and $\hat{S}_e^{a,b,\rho}$ such that the adjoint eigenvalue equation

$$(\mu_1\mu_2)^* \hat{\Psi}_\ell^{a,b,\rho} = \lambda_\ell \hat{\Psi}_\ell^{a,b,\rho}, \quad \ell = 0,1,2,\ldots \tag{4.7}$$

has hypergeometric solutions $\hat{\Psi}_\ell^{a,b,\rho}$ where

$$(\mu_1\mu_2)^* : \hat{S}_e^{a,b,\rho} \to \hat{S}_e^{a,b,\rho}.$$

An evident solution is

$$\hat{\Psi}_\ell^{a,b,\rho}(z) = \Psi_\ell^{a,b,\rho}\left(\frac{q}{b\rho z}\right)$$

$$= \sum_{k=0}^{\ell} \frac{(q^{-2\ell};q^2)_k(\frac{b}{\rho}q^{2\ell-1};q^2)_k}{(q^2;q^2)_k(\frac{a^2q}{b\rho};q^2)_k}\frac{(\frac{aq}{b\rho z};q)_{2k}}{(\frac{q}{\rho z};q)_{2k}}q^{2k} \tag{4.8}$$

with $\hat{S}_e^{a,b,\rho}$ as the space of all finite linear combinations of the $\{\hat{\Psi}_\ell^{a,b,\rho}\}$. The weight function must satisfy the recurrence

$$\frac{w^e_{a,b,\rho}(qz)}{w^e_{a,b,\rho}(z)} = -\frac{(1-az)(1-\frac{1}{\rho z})}{(1-\frac{a}{b\rho z})(1-bz)}. \tag{4.9}$$

This recurrence has many solutions, depending on our choice of the zeros and the poles of w in the z-plane. One of the solutions with the simplest pole structure is

$$w^e_{a,b,\rho}(z) = \frac{(bz;q)_\infty(-\rho z;q)_\infty(-\frac{q}{\rho z};q)_\infty}{(az;q)_\infty(\frac{qa}{\rho bz};q)_\infty(\rho z;q)_\infty} \tag{4.10}$$

where we assume

$$|q| < |\rho| < 1, \qquad |qa| < |\rho b|. \tag{4.11}$$

For C we take the unit circle: $|z| = 1$. We will adopt solution (4.10) in the computations to follow.

Note that

$$\mu_2\Psi_\ell^{a,b,\rho} = \frac{q^2(a-b)(1-q^{-2\ell})(1-\frac{b}{\rho}q^{2\ell-1})}{(1-\frac{a^2q}{b\rho})}\Theta_{\ell-1}^{a,b,\rho} \tag{4.12}$$

where

$$\Theta_{-1}^{a,b,\rho} = 0,$$

$$\Theta_\ell^{a,b,\rho} = \sum_{k=0}^{\ell} \frac{(q^{-2\ell};q^2)_k(\frac{b}{\rho}q^{2\ell+3};q^2)_k}{(q^2;q^2)_k(\frac{a^2q^3}{b\rho};q^2)_k}\frac{(az;q)_{2k+1}q^{2k}}{(bzq;q)_{2k+1}}, \tag{4.13}$$

$$\ell = 0, 1, \ldots.$$

It follows from (4.3)-(4.5) that

$$\mu_1\Theta_{\ell-1}^{a,b,\rho} = \frac{(a^2q-\rho b)}{qa(a-b)}\Psi_\ell^{a,b,\rho}, \quad \ell = 1, 2, \ldots. \tag{4.14}$$

Let $S_o^{a,b,\rho}$ be the space of all finite linear combinations of the $\{\Theta_\ell^{a,b,\rho}\}$. We have the interpretation

$$\mu_1 : S_o^{a,b,\rho} \to S_e^{a,b,\rho}, \ \mu_2 : S_e^{a,b,\rho} \to S_o^{a,b,\rho}.$$

Furthermore, $(\mu_1\mu_2)^*$ factors as $(\mu_1\mu_2)^* = \mu_2^*\mu_1^*$ where

$$\mu_1^* = \frac{q^2}{\rho^2 z}(1-\rho z)\left(1-\frac{\rho z}{q}\right)(I_z - E_z^1),$$

$$\mu_2^* = \frac{b^2q^2}{qa^2}\frac{(1-az)}{(1-\frac{b\rho z}{a})}E_z^1 + \frac{b\rho^2}{qa}\frac{(1-\frac{bz}{q})}{(1-\frac{\rho z}{q})}I_z$$

and

$$\mu_1^*\hat{\Psi}_\ell^{a,b,\rho} = \frac{\frac{q^3}{\rho}(\frac{a}{b}-1)}{(1-\frac{a^2q}{b\rho})}(1-q^{-2\ell})\left(1-\frac{b}{\rho}q^{2\ell-1}\right)\hat{\Theta}_{\ell-1}^{a,b,\rho},$$

$$\mu_2^*\hat{\Theta}_{\ell-1}^{a,b,\rho} = -\frac{b^2\rho^2(1-\frac{a^2q}{b\rho})}{aq^2(a-b)}\hat{\Psi}_\ell^{a,b,\rho}, \tag{4.15}$$

with

$$\hat{\Theta}_{-1}^{a,b,\rho}(z) = 0,$$

$$\hat{\Theta}_\ell^{a,b,\rho}(z) = \sum_{k=0}^{\ell}\frac{(q^{-2\ell};q^2)_k(\frac{b}{\rho}q^{2\ell+3};q^2)_k}{(q^2;q^2)_k(\frac{a^2q^3}{b\rho};q^2)_k}\frac{(\frac{aq}{b\rho z};q)_{2k+1}}{(\frac{q^2}{\rho z};q)_{2k+1}}q^{2k}, \tag{4.16}$$

$$\ell = 0, 1, \ldots.$$

Let $\hat{S}_o^{a,b,\rho}$ be the space of all finite linear combinations of the functions $\{\hat{\Theta}_\ell^{a,b,\rho}\}$. Then we have the interpretations

$$\mu_1^* : \hat{S}_e^{a,b,\ell} \to \hat{S}_o^{a,b,\rho}, \quad \mu_2^* : \hat{S}_o^{a,b,\rho} \to \hat{S}_e^{a,b,\rho}.$$

We now try to determine a weight function $w_{a,b,\rho}^o(z)$ such that the adjoint relation

$$< \mu_1^* f, g >_{a,b,\rho}^o = < f, \mu_1 g >_{a,b,\rho}^e \tag{4.17}$$

holds for all $f \in \hat{S}_e^{a,b,\rho}$, $g \in S_o^{a,b,\rho}$, where

$$< g_1, g_2 >_{a,b,\rho}^o = \oint_C g_1(z)g_2(z)w_{a,b,\rho}^o(z)\frac{dz}{z}. \tag{4.18}$$

A straightforward computation yields

$$w_{a,b,\rho}^o(z) = \frac{\rho^2 bz w_{a,b,\rho}^e(z)}{aq^2(1-bz)(1-\frac{\rho z}{q})}$$

$$= \frac{\rho^2 bz(bqz;q)_\infty(-\rho z;q)_\infty(-\frac{q}{\rho z};q)_\infty}{aq^2(az;q)_\infty(\frac{qa}{\rho bz};q)_\infty(\frac{\rho z}{q};q)_\infty}. \tag{4.19}$$

We can similarly verify that the adjoint relation

$$< \mu_2^* g, f >_{a,b,\rho}^e = < g, \mu_2 f >_{a,b,\rho}^o \tag{4.20}$$

holds for all $f \in S_e^{a,b,\rho}$, $g \in \hat{S}_o^{a,b,\rho}$.

It follows immediately from these adjoint relations and the eigenvalue equations (4.3) and (4.7) that the biorthogonality relations

$$< \hat{\Psi}_\ell^{a,b,\rho}, \Psi_{\ell'}^{a,b,\rho} >_{a,b,\rho}^e = 0, \tag{4.21A}$$

$$< \hat{\Theta}_\ell^{a,b,\rho}, \Theta_{\ell'}^{a,b,\rho} >_{a,b,\rho}^o = 0, \tag{4.21B}$$

hold for all $\ell \neq \ell'$. Our remaining problem is to compute the left-hand sides of expressions (4.21A), (4.21B) for $\ell = \ell'$.

One appropriate operator for this problem is

$$\xi = \frac{q}{z}(1-bqz)\left(1-\frac{b^2\rho z}{a^2}\right)I_z - \frac{q}{z}\frac{(1-bz)(1-bqz)(1-\frac{b\rho z}{aq})}{(1-\frac{az}{q})}E_z^{-1} \tag{4.22}$$

which satisfies

$$\xi p_{2k+1}^{a,bq} = (a-bq)\frac{b\rho}{a^2}\left(1-\frac{q^{2k+1}a^2}{b\rho}\right)p_{2k}^{a,bq^2},$$

$$k = 0,1,2,\dots \tag{4.23}$$

where the basis functions $p_n^{a,b}$ are defined by (4.1). It follows that

$$\xi\Theta_\ell^{a,b,\rho} = -q(a-bq)\left(1-\frac{b\rho}{a^2q}\right)\Psi_\ell^{a,bq^2,\rho q^{-2}}, \tag{4.24}$$

$$\ell = 0,1,\dots$$

and that ξ has the interpretation

$$\xi : S_o^{a,b,\rho} \to S_e^{a,bq^2,\rho q^{-2}}.$$

We define the adjoint operator

$$\xi^* : \hat{S}_e^{a,bq^2,\rho q^{-2}} \to \hat{S}_o^{a,b,\rho}$$

by

$$< g, \xi f >_{a,bq^2,\rho q^{-2}}^{e} = < \xi^* g, f >_{a,b,\rho}^{o} \tag{4.25}$$

where $g \in \hat{S}_e^{a,bq^2,\rho q^{-2}}$, $f \in S_o^{a,b,\rho}$. A straightforward computation yields

$$\xi^* = qE_z^1 + \frac{a}{b}\frac{(1-\frac{b^2\rho z}{a^2})}{(1-\frac{\rho z}{q^2})}I_z \tag{4.26}$$

and

$$\frac{(a-bq)}{q}\xi^*\hat{\Psi}_\ell^{a,bq^2,\rho q^{-2}} = -\frac{q^2}{a}\frac{(1-\frac{a^2q^{2\ell+1}}{b\rho})}{(1-\frac{a^2q}{b\rho})}(b^2-a^2q^{-2\ell-2})\hat{\Theta}_\ell^{a,b,\rho}. \tag{4.27}$$

Set

$$\|\Psi_\ell^{a,b,\rho}\|_e^2 = < \hat{\Psi}_\ell^{a,b,\rho}, \Psi_\ell^{a,b,\rho} >_{a,b,\rho}^{e}, \quad \|\Theta_\ell^{a,b,\rho}\|_o^2 = < \hat{\Theta}_\ell^{a,b,\rho}, \Theta_\ell^{a,b,\rho} >_{a,b,\rho}^{o}, \tag{4.28}$$

$$\ell = 0,1,2,\dots.$$

From (4.14), (4.15), and (4.17) with $f = \hat{\Psi}_\ell^{a,b,\rho}$, $g = \Theta_{\ell-1}^{a,b,\rho}$ we have

$$\|\Psi_\ell^{a,b,\rho}\|_e^2 = -\frac{a\rho q^2(a-b)^2(1-q^{-2\ell})(1-\frac{b}{\rho}q^{2\ell-1})}{(b\rho-a^2q)^2}\|\Psi_{\ell-1}^{a,b,\rho}\|_o^2. \tag{4.29}$$

Relation (4.20) yields the same recurrence. Expressions (4.24), (4.25), and (4.27) with $f = \Theta_\ell^{a,b,\rho}$, $g = \hat{\Psi}_\ell^{a,bq^2,\rho q^{-2}}$ produce

$$\|\Psi_\ell^{a,bq^2,\rho q^{-2}}\|_e^2 = -\frac{a^2q^3(1-\frac{a^2q^{2\ell+1}}{b\rho})(b^2-a^2q^{-2\ell-2})}{b\rho(a-bq)^2(1-\frac{a^2q}{b\rho})^2}\|\Theta_\ell^{a,b,\rho}\|_o^2. \tag{4.30}$$

It follows from these results that if we know $\|1^{a,b,\rho}\|_e^2$ for all a, b, ρ then we can compute recursively expressions (4.28) for all ℓ. In general these norms will be nonzero.

Replacing ρz by z in (4.6), (4.10) we have

$$\|1^{a,b,\rho}\|_e^2 = G(u,v,q) = \frac{1}{2\pi}\oint_C \frac{(\frac{uz}{v};q)_\infty(-z;q)_\infty(-\frac{q}{z};q)_\infty}{(uz;q)_\infty(\frac{qv}{z};q)_\infty(z;q)_\infty}\frac{dz}{z} \tag{4.31}$$

where $u = a/\rho$, $v = a/b$. From the relation

$$< \hat{\Psi}_0^{a,b,\rho}, \Psi_1^{a,b,\rho} >_e = 0$$

it is straightforward to compute the recurrence

$$\|1^{aq^2,bq^2,\rho}\|_e^2 = \frac{(1 - \frac{a^2q}{b\rho})}{(1 - \frac{bq}{\rho})}\|1^{a,b,\rho}\|_e^2$$

or $G(q^2u, v, q) = (1 - uvq)G(u, v, q)/(1 - uq/v)$, which implies

$$G(u, v, q) = \frac{(\frac{uq}{v}; q^2)_\infty}{(uvq; q^2)_\infty}\mathscr{G}(v, q). \tag{4.32}$$

To finish the computation of (4.31) we utilize the recurrence operator

$$\eta = \frac{(aq - b)}{(1 - az)}\left[\left(1 - \frac{bz}{q}\right)E_z^{-1} + \frac{b}{aq}\left(1 - \frac{a^2qz}{b}\right)I_z\right]. \tag{4.33}$$

Here,

$$\eta p_n^{a,b} = aq(1 - q^{-n})p_{n-2}^{aq,b} + \frac{(-b^2 + a^2q^{2-n})}{aq}p_n^{aq,b},$$

hence

$$\eta\Psi_\ell^{a,b,\rho} = -\frac{q^{-2\ell}(b - \frac{a^2}{\rho}q^{2\ell+1})}{abq(1 - \frac{a^2q}{b\rho})}(-a^2q^2 + b^2q^{2\ell})\Psi_\ell^{aq,b,\rho}. \tag{4.34}$$

Interpreting $\eta : S_e^{a,b,\rho} \to S_e^{aq,b,\rho}$ we compute the adjoint $\eta^* : \hat{S}_e^{aq,b,\rho} \to \hat{S}_e^{a,b,\rho}$ with the result

$$\eta^* = (aq - b)\left[\frac{(1 - \rho z)}{\rho z}(1 - az)E_z^1 + \frac{b}{aq}\left(1 - \frac{qa}{\rho bz}\right)\left(1 - \frac{a^2qz}{b}\right)I_z\right]. \tag{4.35}$$

This leads to the recurrence

$$\eta^*\hat{\Psi}_\ell^{aq,b,\rho} = -\frac{(aq - b)^2}{aq}\left(1 - \frac{a^2q}{b\rho}\right)\hat{\Psi}_\ell^{a,b,\rho}. \tag{4.36}$$

Thus the relation

$$< \eta^*1, 1 >_{a,b,\rho}^e = < 1, \eta 1 >_{aq,b,\rho}^e$$

implies

$$\|1^{aq,b,\rho}\|_e^2 = \left(1 - \frac{a^2q}{bm}\right)\left(\frac{b - aq}{b + aq}\right)\|1^{a,b,\rho}\|_e^2$$

or $\mathscr{G}(qv, q) = (1-qv)\mathscr{G}(v, q)/(1+qv)$. We conclude $\mathscr{G}(v, q) = (-vq; q)_\infty\mathscr{K}(q)/(vq; q)_\infty$ where $\mathscr{K}(q)$ is to be determined. Setting $u = 0$, $v = 1$ in (4.31) we find $G(0, -1, q) = 1 = (q; q)_\infty\mathscr{K}(q)/(-q; q)_\infty$, so

$$\|1^{a,b,\rho}\|_e^2 = \frac{(-q; q)_\infty(\frac{bq}{\rho}; q^2)_\infty(-\frac{aq}{b}; q)_\infty}{(q; q)_\infty(\frac{a^2q}{b\rho}; q^2)_\infty(\frac{aq}{b}; q)_\infty}. \tag{4.37}$$

Expanding the original integral $\|1^{a,b,\rho}\|_e^2$ by residues inside the circle C, we see that our result is equivalent to the summation formula

$${}_2\varphi_1\left(\begin{matrix} vq, & uvq \\ uq & \end{matrix}; -\frac{1}{v}\right) = \frac{(-q; q)_\infty(\frac{uq}{v}; q^2)_\infty(uvq^2; q^2)_\infty}{(uq; q)_\infty(-\frac{1}{v}; q)_\infty}. \tag{4.38}$$

This is a q-analogue of Kummer's Theorem, first proved by Andrews (1973); see also Andrews (1977, p. 20).

Acknowledgments. Willard Miller wishes to thank Dick Askey and Dennis Stanton for consultation on this work. We also thank Dennis Stanton for pointing out the fact in (2.19).

Note added in proof. A referee points out that the system of biorthogonal functions in §4 is a special case of a system found by J.A. Wilson in 1977 but not yet published.

REFERENCES

[1] A. K. AGARWAL, E. G. KALNINS, AND W. MILLER, *Canonical equations and symmetry techniques for q-series*, SIAM J. Math. Anal., 18 (1987), pp. 1519–1538.
[2] G. E. ANDREWS, *On the q-analogue of Kummer's theorem and applications*, Duke Math. J., 40 (1973), pp. 525–528.
[3] G. E. ANDREWS, *The Theory of Partitions*, Addison-Wesley, Reading, MA, 1977.
[4] G. E. ANDREWS AND R. ASKEY, *Classical orthogonal polynomials*, Lecture Notes # 1171, Springer-Verlag, New York, Berlin, 1985, pp. 36–62.
[5] R. ASKEY AND R. ROY, *More q-beta integrals*, Rocky Mtn. J. Math., 26 (1986), pp. 365–372.
[6] R. ASKEY AND J. WILSON, *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*, Memoirs of the AMS, 319 (1985).
[7] W. N. BAILEY, *Generalized hypergeometric series*, Cambridge University Press, 1935, Reprinted by Stechert-Hafner, New York, 1964.
[8] M. E.-H. ISMAIL AND J. A. WILSON, *Asymptotic and generating relations for the q-Jacobi and $4\varphi_3$ polynomials*, J. Approx. Theory, 36 (1982), pp. 43–54
[9] E. G. KALNINS AND W. MILLER, *Symmetry techniques for q-series: Askey-Wilson polynomials*, Rocky Mtn. J. Math., to appear
[10] W. MILLER, *A note on Wilson polynomials*, SIAM J. Math. Anal., 18 (1987), pp. 1221–1226.
[11] A. F. NIKIFOROV AND S. K. SUSLOV, *Classical orthogonal polynomials of a discrete variable on nonuniform lattices*, Lett. Math. Phys., 11 (1986), pp. 27–34
[12] A. F. NIKIFOROV, S. K. SUSLOV, AND V. B. UVAROV, *Classical Orthogonal Polynomials of a Discrete Variable*, Nauka, Moscow (in Russian), 1985.
[13] L. J. SLATER, *Generalized Hypergeometric Functions*, Cambridge University Press, Cambridge, 1966.
[14] G. N. WATSON, *The continuation of functions defined by generalized hypergeometric series*, Trans.Cambridge Phil. Soc., 21 (1910), pp. 281–299
[15] E. T. WHITTAKER AND G. N. WATSON, *A Course in Modern Analysis*, Cambridge University Press, Cambridge, 1958 .
[16] J. WILSON, *Some hypergeometric orthogonal polynomials*, SIAM J. Math. Anal., 11 (1980), pp. 690–701

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

019

UNIFORM ASYMPTOTIC EXPANSIONS OF LAGUERRE POLYNOMIALS*

C. L. FRENZEN† AND R. WONG‡

Abstract. Two asymptotic expansions are obtained for the Laguerre polynomial $L_n^{(\alpha)}(x)$ for large n and fixed $\alpha > -1$. These expansions are uniformly valid in two overlapping intervals covering the entire x-axis. The leading terms of both agree with the two asymptotic formulas given by Erdélyi who used the theory of differential equations. Our approach is based on two integral representations for the Laguerre polynomials. The phase function of one of these integrals has two coalescing saddle points, and to this one the cubic transformation introduced by Chester, Friedman, and Ursell is applied. The phase function of the other integral also has two coalescing saddle points, but in addition it has a simple pole. Moreover, the saddle points coalesce onto this pole. In this case a rational transformation is used, which mimics the singular behavior of the phase function. In both cases explicit expressions are given for the remainders associated with the asymptotic expansions.

Key words. uniform asymptotic expansion, Laguerre polynomial, Airy function, Bessel function

AMS(MOS) subject classifications. primary 41A60, 33A65

1. Introduction.

Many special functions of mathematical physics have integral representations of the form

$$I(\lambda) = \int_C g(z)\, e^{\lambda f(z,t)}\, dz, \tag{1.1}$$

where C is a contour in the complex plane, λ is a large positive parameter, and $f(z, t)$ and $g(z)$ are analytic functions of their arguments. For fixed t the asymptotic expansion of $I(\lambda)$ can often be found by the method of steepest descents, which shows that the major contributions to $I(\lambda)$ come from the saddle points of $f(z, t)$, i.e., the points where $\partial f(z, t)/\partial z$ vanishes. In general, the saddle points depend on t. As t varies continuously, the saddle points may coalesce with each other, and the form of the asymptotic expansion may change. The problem at hand is to obtain an asymptotic expansion for large λ that is uniform with respect to t as t ranges over a given connected set that is not necessarily bounded. When $\partial f(z, t)/\partial z$ has exactly two simple zeros that coalesce into a double zero, Chester, Friedman, and Ursell suggested, in their innovative paper [2], the use of the cubic transformation

$$f(z, t) = \tfrac{1}{3} u^3 - \zeta(t) u + \eta(t), \tag{1.2}$$

where $\zeta(t)$ and $\eta(t)$ are determined explicitly from the requirement that the transformation (1.2) from z to u be analytic in a neighborhood containing the two saddle points. As a result, a uniform asymptotic expansion of the form

$$\exp\{-\lambda\eta(t)\}\left[\frac{\mathrm{Ai}\,(\lambda^{2/3}\zeta)}{\lambda^{1/3}} \sum_{s=0}^{\infty} \frac{A_s(\alpha)}{\lambda^s} + \frac{\mathrm{Ai}'\,(\lambda^{2/3}\zeta)}{\lambda^{2/3}} \sum_{s=0}^{\infty} \frac{B_s(\alpha)}{\lambda^s}\right] \tag{1.3}$$

was obtained, where Ai and Ai′ are the Airy function and its derivative, respectively. Airy function expansions of the form (1.3) were in fact first found by Langer [7] and

* Received by the editors January 12, 1987; accepted for publication (in revised form) December 21, 1987.
† Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.
‡ Department of Applied Mathematics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2.
The work of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A7359.

Olver [11] in their study of asymptotic solutions to the differential equation

$$\frac{d^2w}{dz^2}=\{u^2p(z)+q(z)\}w \tag{1.4}$$

for large values of u. When the function $p(z)$ has only a simple turning point, (1.4) can be transformed into one which is approximately the same as the equation satisfied by the Airy functions. For an elegant version of this theory, see the book by Olver [13, Chap. 11].

In this paper we shall study the behavior of the integral (1.1), when $f(z, t)$ is given by

$$f(z, t)=z-t\coth z. \tag{1.5}$$

Note that this function has two symmetrically located saddle points

$$z_{\pm}=\pm i\sin^{-1}\sqrt{t}$$

and, in addition, a simple pole at $z=0$. As t tends to zero, the saddle points coalesce with each other and also with the pole. The simplest function which also possesses these essential features is provided by the rational function $u-A^2(t)/u$. Thus, instead of (1.2), the appropriate transformation is

$$z-t\coth z=u-\frac{A^2(t)}{u}, \tag{1.6}$$

where $A(t)$ is to be determined. As a consequence of (1.6), we obtain a uniform asymptotic expansion of $I(\lambda)$ of the form

$$\frac{J_\alpha(2\lambda A)}{A^\alpha}\sum_{k=0}^{\infty}\frac{C_{2k}(t)}{\lambda^{2k}}-\frac{J_{\alpha+1}(2\lambda A)}{A^{\alpha+1}}\sum_{k=0}^{\infty}\frac{D_{2k+1}(t)}{\lambda^{2k+1}}, \tag{1.7}$$

where J_α is the Bessel function of the first kind. A transformation similar to (1.6) has been suggested by Temme [15, eq. (6.12)], but, as he points out, his argument is only formal (cf. [15, lines 6–8, p. 313]). It is of interest to observe that expansions of this form have also been given in the theory of differential equations. They represent the asymptotic solutions to (1.4), when the coefficient function $p(z)$ has a simple pole (see [12] and also [13, Chap. 12]).

The integral (1.1) with $f(z, t)$ given by (1.5) arises in the study of the asymptotic behavior of the Laguerre polynomial $L_n^{(\alpha)}(x)$. The best known result in this area is probably that of Erdélyi [4]. Erdélyi put $\nu=4N=4n+2\alpha+2$, and gave two uniform asymptotic formulas for $L_n^{(\alpha)}(\nu t)$, as $n\to\infty$, where α is fixed and nonnegative, and t is real. One formula holds uniformly for $t\leqq a$ and the other for $t\geqq b$, where a and b are two fixed numbers, $0<b<a<1$. Note that these two intervals overlap and between them cover the entire real axis. Erdélyi's formulas were considered important results by Szegö [14, p. 243], and their validity was extended from $\alpha\geqq 0$ to $\alpha>-1$ by Muckenhoupt [10] through the use of recurrence relations. It should be mentioned that since Laguerre polynomials can be expressed in terms of confluent hypergeometric functions [13, p. 259], it is possible to derive their infinite asymptotic expansions with error bounds directly from those for Whittaker functions given by Olver [13, pp. 412, 446–447]. However, so far, all of these results have only been obtained from the theory of differential equations.

The object of the present paper is to obtain the same kind of asymptotic expansions for the Laguerre polynomial from its definite integral representations. An attempt in this direction was made earlier by Wyman [18], but the range of validity of his result is much more restricted than that of Erdélyi. The method we shall use depends heavily on establishing that the transformation (1.6) has one branch $u = u(z, t)$ that is analytic in z and continuous in t, and that the correspondence $u \leftrightarrow z$ is one-to-one. This can be done in two ways. It can either be proved directly, as was done by Copson [3, Chap. 10] and Olver [13, Chap. 9, § 12], who both used the cubic transformation for Bessel and Anger functions, respectively, or it can be obtained as a consequence of a general theorem corresponding to the one given by Chester, Friedman, and Ursell [2, Thm. 1]. Although it is possible to establish such a general result like the one for the cubic transformation in [2], it will hold only locally and not globally. We therefore find it preferable to adopt the approach taken by Copson and Olver. Indeed, it will be shown that the mapping $z \leftrightarrow u$ in (1.6) is actually one-to-one and analytic along an infinite loop starting at $-\infty$, enclosing the origin, and ending at $-\infty$.

Here we wish to remark that although we consider only the case of Laguerre polynomials, the method presented in this paper is quite general. It can be applied to many other integrals whose phase function $f(z, t)$ has two symmetrically located saddle points and a simple pole. Since there are functions that are expressible in the form of an integral but do not satisfy any second-order linear differential equation, a complete theory of asymptotic analysis requires integral as well as differential equation methods for deriving uniform asymptotic expansions.

The present paper is arranged as follows. In § 2 we first show that the Laguerre polynomial $L_n^{(\alpha)}(\nu t)$ has an integral representation whose phase function is given by (1.5), and then we use the rational transformation (1.6) to reduce it to canonical form. For $-\infty < t \leqq a < 1$, the one-to-one nature of the transformation (1.6) is demonstrated in § 3. The derivation of the asymptotic expansion is given in § 4. In § 5, we study the case $0 < b \leqq t < \infty$, and obtain an Airy function expansion. A concluding remark is given in § 6.

2. Reduction to a canonical integral. We start with the integral representation [14, p. 384]

$$e^{-x/2}L_n^{(\alpha)}(x) = \frac{1}{2\pi i}\int_{-\infty}^{(0+)} \exp\left\{-\frac{x}{2}\frac{1+e^{-z}}{1-e^{-z}}\right\}(1-e^{-z})^{-\alpha-1}\, e^{nz}\, dz, \tag{2.1}$$

where the path of integration is the usual loop which begins and ends at $-\infty$ and encircles the origin in the positive direction. (Note that this is not the integral used by Wyman.) *Throughout this paper we shall assume that* $\alpha > -1$. Clearly (2.1) can be written as

$$\begin{aligned} e^{-x/2}L_n^{(\alpha)}(x) = \frac{1}{2\pi i}\int_{-\infty}^{(0+)} &\exp\left\{-\frac{x}{2}\coth\frac{z}{2} + Nz\right\} \\ &\cdot\left(\frac{\sinh z/2}{z/2}\right)^{-\alpha-1} z^{-\alpha-1}\, dz \end{aligned} \tag{2.2}$$

with $N = n + \frac{1}{2}(\alpha+1)$. If we replace z by $2z$ and let $\nu = 4N$ and $x = \nu t$, $-\infty < t < 1$, then (2.2) becomes

$$e^{-\nu t/2}L_n^{(\alpha)}(\nu t) = \frac{2^{-\alpha}}{2\pi i}\int_{-\infty}^{(0+)} \exp\left\{\frac{\nu}{2}f(z, t)\right\}\left(\frac{\sinh z}{z}\right)^{-\alpha-1} z^{-\alpha-1}\, dz, \tag{2.3}$$

where $f(z, t)$ is given in (1.5):

$$f(z, t) = z - t \coth z. \tag{2.4}$$

We shall represent $f(z, t)$ by the rational function

$$f(z, t) = u - \frac{A^2(t)}{u}. \tag{2.5}$$

For this to be an analytic transformation in the regions of interest, we must have $dz/du \neq 0$ or ∞. Now

$$f_z(z, t)\frac{dz}{du} = 1 + \frac{A^2(t)}{u^2} \tag{2.6}$$

and $f_z(z, t)$ vanishes at $z = z_+$ and z_- in $|\operatorname{Im} z| < \pi/2$, where

$$z_\pm = \begin{cases} \pm i \sin^{-1}\sqrt{t}, & 0 \leqq t < 1, \\ \mp \sinh^{-1}\sqrt{-t}, & t < 0. \end{cases}$$

Since the right-hand side of (2.6) vanishes at $u = \pm iA(t)$, we must make $z = z_+$ correspond to $u = +iA(t)$, and $z = z_-$ to $u = -iA(t)$. This gives

$$A(t) = \begin{cases} \frac{1}{2}[\sin^{-1}\sqrt{t} + \sqrt{t(1-t)}], & 0 \leqq t < 1, \\ \frac{i}{2}[\sinh^{-1}\sqrt{-t} + \sqrt{t(t-1)}], & t < 0. \end{cases} \tag{2.7}$$

Note that our $A(t)$ is exactly the same as Erdélyi's $\psi(t)$. In § 3, it will be shown that with this choice, the transformation (2.5) is one-to-one and analytic along the whole infinite loop given in (2.3). Thus, changing the variable to u, we obtain

$$e^{-\nu t/2} L_n^{(\alpha)}(\nu t) = \frac{2^{-\alpha}}{2\pi i} \int_{-\infty}^{(0+)} u^{-\alpha-1} h(u) \exp\left\{\frac{\nu}{2}\left(u - \frac{A^2(t)}{u}\right)\right\} du, \tag{2.8}$$

where

$$h(u) = \left(\frac{\sinh z(u)}{z(u)}\right)^{-\alpha-1} \left(\frac{z(u)}{u}\right)^{-\alpha-1} \frac{dz}{du}. \tag{2.9}$$

Here we have assumed that the mapping $z \leftrightarrow u$ preserves the loop nature of the path of integration. Note that the function h in (2.9) depends also on the variable t; thus, $h = h(u, t)$. However, for simplicity, we shall not indicate this dependence explicitly.

3. The transformation (1.6). The properties of the mapping between z and u are best seen by introducing an intermediate variable Z defined by

$$z - t \coth z = Z = u - \frac{A^2(t)}{u}. \tag{3.1}$$

For our purpose, it suffices to consider only the strip $|\operatorname{Im} z| \leqq \pi/2$. We first restrict ourselves to the case $0 < t < 1$. The half-strip $\{z: \operatorname{Re} z \leqq 0, 0 \leqq \operatorname{Im} z \leqq \pi/2\}$ is shown in Fig. 1. Its image in the Z-plane is depicted in Fig. 2. Note that as z traverses once along the indented boundary $ABCC'DEFA$ in Fig. 1, Z also traverses exactly once along the corresponding curve in Fig. 2. Here we treat the straight lines $C'D$ and DE in Fig. 2 as distinct parts of the boundary (see [17, p. 375, lines 22–25]). Hence, by

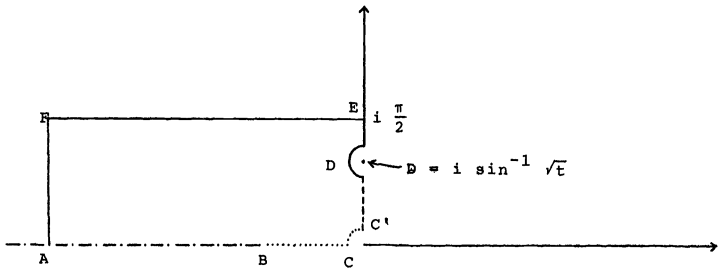


FIG. 1. *z-plane* $(0<t<1)$.

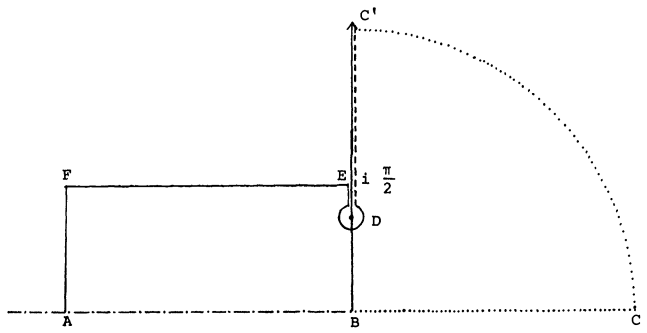


FIG. 2. *Z-plane.*

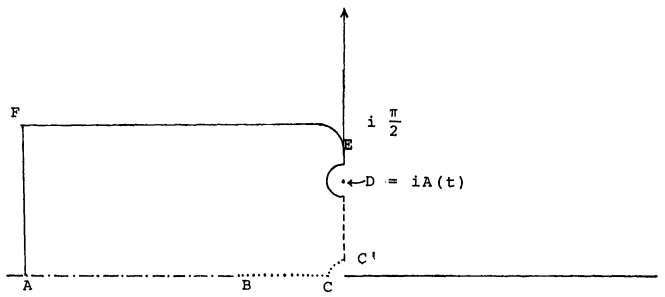


FIG. 3. *u-plane.*

Theorem 4.5 in [9, vol. 2, p. 118], $\phi = z - t\coth z$ is one-to-one in the interior of the region bounded by this curve (see also [16, §§ 6.45 and 6.46]).

We next consider the mapping $\psi: u \to Z$ defined by $\psi(u) = u - A^2(t)/u$. By the same argument as above, when u traverses once along the boundary of the region $ABCC'DEFA$ in Fig. 3, Z goes once around the corresponding curve in the Z-plane. Hence ψ is one-to-one in the interior of the region $ABCC'DEFA$ in Fig. 3. The equation of the boundary curve EF in Fig. 3 is given implicitly by

$$\frac{\pi}{2} = s + \frac{A^2(t)s}{r^2+s^2}, \qquad \operatorname{Re} Z = r - \frac{A^2(t)r}{r^2+s^2},$$

where $u = r + is$ and $r < 0$. Since $r \leqq 0$ and $\operatorname{Re} Z \leqq 0$, we have, from the second equation above, $A^2(t) \leqq r^2+s^2$. This together with the first equation implies that $s \geqq \pi/4$, i.e., the curve EF in Fig. 3 remains in the region $\operatorname{Im} u \geqq \pi/4$. Also, from (2.7), we have $A(0) = 0$, $A'(t) > 0$ for $0 < t < 1$ and $A(1^-) = \pi/4$. Hence, $0 < A(t) < \pi/4$ for $0 < t < 1$. Therefore, for $0 < t \leqq a < 1$, the point D in Fig. 3 is at a positive distance away from the curve EF.

The transformation $z \leftrightarrow u$ is obtained by composing $\phi^{-1}: Z \to z$ and $\psi: u \to Z$. Since the transformations $z \leftrightarrow Z$ and $u \leftrightarrow Z$ are one-to-one within the boundary $ABCC'DEFA$, so is $z \leftrightarrow u$. Let $z = x + iy$ and $u = r + is$. It can be shown by direct computation that the real parts of $z - t \coth z$ and $u - (1/u)A^2(t)$ are odd in x and r, respectively, and even in y and s, respectively. Similarly, the imaginary parts of these functions are odd in y and s, respectively, and even in x and r, respectively. Hence, the mapping of the rest of the strip $|\text{Im } z| \leqq \pi/2$ is deducible from Figs. 1 and 3 by reflection in the real and imaginary axes. This establishes the one-to-one and analytic nature of the function $u(z, t)$ in $|\text{Im } z| \leqq \pi/2$, except possibly at $z = z_{\pm}$ and $z = 0$. From the above argument (cf. Figs. 1 and 3), it is also evident that neighborhoods of these points are mapped into neighborhoods of their corresponding images. Consequently, $u(z, t)$ is bounded and analytic at these points. (Note also that near $z = 0$ and $t = 0$, we have $u \sim (1/t)A^2(t)z$ and $A^2(t) \sim t$, respectively.)

To emphasize what has been proved, we state again that the mapping $z \leftrightarrow u$, when $0 < t < 1$, is one-to-one and analytic from $|\text{Im } z| \leqq \pi/2$ to its image in the u-plane. The same properties, when $-\infty < t < 0$, can be established in a similar manner. In the latter case, the regions bounded by $ABCC'DEFA$ in the z-, Z-, and u-planes are shown in Figs. 1′, 2′, and 3′, respectively. Arguments similar to ours have been used earlier by Copson [3, § 49] and Olver [13, Chap. 9, § 12.3].

We have therefore proved (2.8) for the cases $0 < t < 1$ and $-\infty < t < 0$. The fact that (2.8) holds also at $t = 0$ follows from continuity. To show that $u(z, t)$ is continuous in t, we note that when $z = 0$, u is identically zero and hence is continuous in t. For $z \neq 0$, we have $u \neq 0$, as there is a one-to-one correspondence between z and u. Thus, for $z \neq 0$, (3.1) is equivalent to the quadratic equation $u^2 - (z - t \coth z)u - A^2(t) = 0$. Since $A^2(t)$ is continuous in t and the solutions of the quadratic equation depend continuously on its coefficients, u is continuous in t for $-\infty < t < 1$.

It may be of interest to note that, by using Hartogs' theorem [6, pp. 27-28] in the theory of several complex variables, we can actually prove that $u(z, t)$ is analytic in both variables in a neighborhood of $z = t = 0$. However, no use is made of this property in our later analysis.

4. Derivation of the expansion. We now return to the integral in (2.8) and deform the loop path of integration so that it consists of two straight lines along the negative real axis and a circle centered at the origin with radius ρ. The radius will be specified later, and depends on whether t is close to, or bounded away from, the origin. From (3.1) it can be shown that $z(u)$ is an odd function of u. Thus, (2.9) implies that $h(u)$ is an even function. Put $h_0(u) = h(u)$ and write

$$h_0(u) = \alpha_0 + \frac{\beta_0}{u} + \left(1 + \frac{A^2(t)}{u^2}\right) g_0(u), \tag{4.1}$$

where α_0, β_0, and $g_0(u)$ are to be determined. Since $h(u)$ is even, $h_0(iA) = h_0(-iA)$ and it follows that $\beta_0 = 0$, $\alpha_0 = h_0(iA)$. Thus

$$g_0(u) = u^2 \frac{h_0(u) - h_0(iA)}{u^2 - (iA)^2}. \tag{4.2}$$

Clearly, $g_0(u)$ is analytic for $u^2 \neq (iA)^2$ and has removable singularities at $u = \pm iA$. Near $u = \pm iA$,

$$g_0(u) = \frac{u^2}{u + (\pm iA)} \frac{h_0(u) - h_0(\pm iA)}{u - (\pm iA)},$$

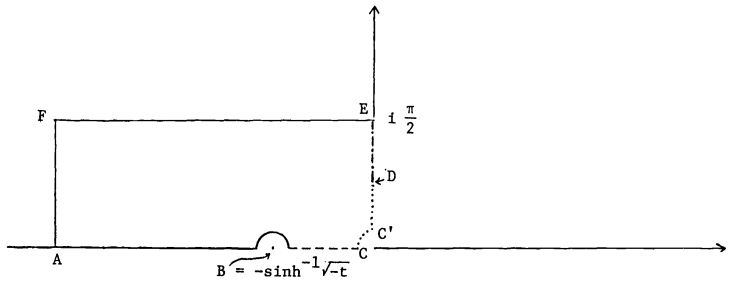


FIG. 1'. *z-plane* ($t<0$).

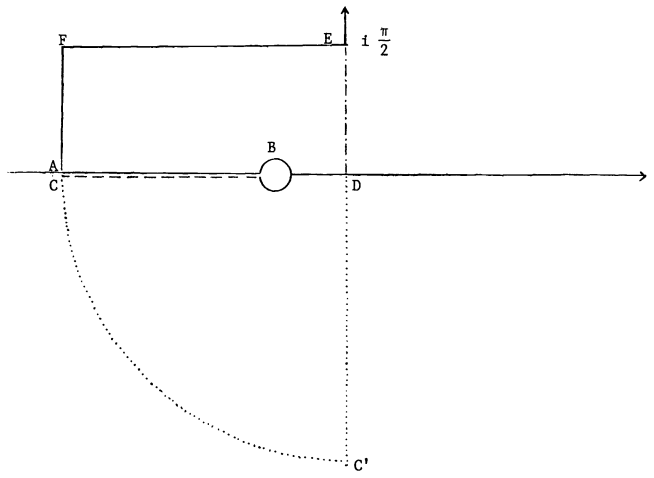


FIG. 2'. *Z-plane* ($t<0$).

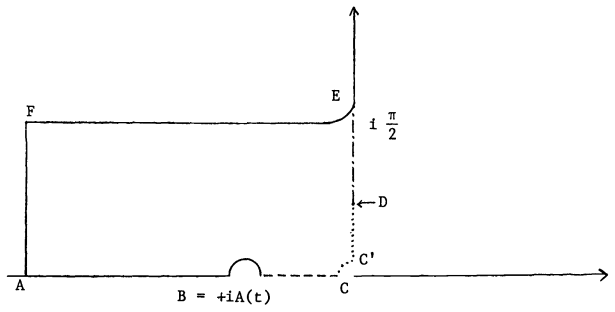


FIG. 3'. *u-plane* ($t<0$).

where either all "+" or all "−" signs are to be taken. From this, we conclude that $g_0(u)$ is analytic everywhere in the domain of $h_0(u)$. Now substitute (4.1) in (2.8), and express the first integral in terms of the Bessel function J_α. To the second integral, we apply an integration by parts. The integrated term vanishes, since $z(u)\sim u$ and $dz/du = O(1)$ as $u\to-\infty$, for fixed t. The final result is

$$2^\alpha e^{-\nu t/2}L_n^{(\alpha)}(\nu t)=\frac{\alpha_0}{A^\alpha}J_\alpha(\nu A)-\left(\frac{2}{\nu}\right)\frac{1}{2\pi i}\int_{-\infty}^{(0+)}u^{-\alpha-1}h_1(u)\cdot\exp\left\{\frac{\nu}{2}\left(u-\frac{A^2(t)}{u}\right)\right\}du, \tag{4.3}$$

where

$$h_1(u) = g_0'(u) - \frac{\alpha+1}{u} g_0(u).$$

A similar technique of integration by parts has been used by Temme [15, eq. (6.2)].

We now digress to briefly discuss the asymptotic behavior of $z^{(n)}(u)$, for fixed t, as $u \to -\infty$, $n = 0, 1, 2, \cdots$. From (1.6) (or (3.1)), it is easily seen that $z(u) \sim u$ as $u \to -\infty$. Differentiating (1.6) with respect to u, we obtain

$$(1 + t\ \mathrm{csch}^2 z) z'(u) = 1 + \frac{A^2(t)}{u^2}, \tag{4.4}$$

which in turn yields $z'(u) = O(1)$ as $u \to -\infty$. From (4.4), we also have, by Leibniz's rule,

$$\begin{aligned}(1 + t\ \mathrm{csch}^2 z) z^{(n+1)}(u) + \sum_{j=1}^{n} \binom{n}{j-1} (1 + t\ \mathrm{csch}^2 z(u))^{(n-j+1)} z^{(j)}(u) \\ = (-1)^n A^2(t) \frac{(n+1)!}{u^{n+2}}.\end{aligned} \tag{4.5}$$

Since every term in the above sum is exponentially decaying, (4.5) implies, by induction, that for fixed t,

$$\frac{d^{n+1} z}{du^{n+1}} = O\left(\frac{1}{u^{n+2}}\right), \qquad n \geqq 1, \quad \text{as } u \to -\infty. \tag{4.6}$$

To derive the form of the expansion given in (1.7), we repeat the procedures indicated in (4.1) and (4.3). Thus, we define recursively

$$\begin{aligned} h_0(u) &= h(u), \\ h_n(u) &= \alpha_n + \frac{\beta_n}{u} + \left(1 + \frac{A^2}{u^2}\right) g_n(u), \\ h_{n+1}(u) &= g_n'(u) - \frac{\alpha+1}{u} g_n(u). \end{aligned}$$

Using induction, we can show that $h_n(u) = O(1)$, as $u \to -\infty$, for fixed t and for all $n \geqq 0$. From this, it follows that we also have $g_n(u) = O(1)$, as $u \to -\infty$, for fixed t and for $n \geqq 0$. Furthermore, we can show that for $n \geqq 0$, $g_{2n}(u)$ and $h_{2n}(u)$ are even analytic functions, and that $g_{2n+1}(u)$ and $h_{2n+1}(u)$ are odd analytic functions. Consequently, we have

$$\alpha_{2n} = h_{2n}(iA), \qquad \alpha_{2n+1} = 0$$

and

$$\beta_{2n} = 0, \qquad \beta_{2n+1} = iA h_{2n+1}(iA).$$

Thus, for each n, both α_n and β_n are continuous in t for $-\infty < t < 1$. Repeated application of integration by parts then gives

$$\begin{aligned} 2^{\alpha} e^{-\nu t/2} L_n^{(\alpha)}(\nu t) = \frac{J_{\alpha}(\nu A)}{A^{\alpha}} \sum_{k=0}^{[(p-1)/2]} \alpha_{2k} \left(\frac{2}{\nu}\right)^{2k} \\ - \frac{J_{\alpha+1}(\nu A)}{A^{\alpha+1}} \sum_{k=0}^{[p/2]-1} \beta_{2k+1} \left(\frac{2}{\nu}\right)^{2k+1} + \varepsilon_p, \end{aligned} \tag{4.7}$$

where

$$\varepsilon_p = \left(-\frac{2}{\nu}\right)^p \cdot \frac{1}{2\pi i} \int_{-\infty}^{(0+)} u^{-\alpha-1} h_p(u) \exp\left\{\frac{\nu}{2}\left(u - \frac{A^2(t)}{u}\right)\right\} du. \tag{4.8}$$

To estimate ε_p, we recall the auxiliary function

$$\tilde{J}_\alpha(z) = \begin{cases} J_\alpha(z) & \text{if } z \text{ is imaginary or } 0 \leqq z \leqq \delta, \\ (|J_\alpha(z)|^2 + |Y_\alpha(z)|^2)^{1/2} & \text{if } z > \delta, \end{cases} \tag{4.9}$$

introduced by Erdélyi [4], where δ is chosen so that $J_\alpha(z) \neq 0$ when $0 < |z| \leqq \delta$ and $\delta > 0$. Furthermore, we define

$$\tilde{\beta}_p(t) = \begin{cases} 1 & \text{for } -\eta \leqq t \leqq a < 1, \\ |\beta_p| & \text{for } -\infty < t < -\eta, \end{cases} \tag{4.10}$$

and define $\tilde{\alpha}_p(t)$ in a similar manner, where α_p and β_p are the coefficients in the expansion (4.7), and η is a positive number to be chosen later. We shall now show that for $-\infty < t \leqq a < 1$,

$$\begin{aligned} |\varepsilon_p| &\leqq \frac{M_p}{\nu^p} \tilde{\beta}_p(t) \left|\frac{\tilde{J}_{\alpha+1}(\nu A(t))}{A(t)^{\alpha+1}}\right| \quad \text{if } p \text{ is odd}, \\ |\varepsilon_p| &\leqq \frac{N_p}{\nu^p} \tilde{\alpha}_p(t) \left|\frac{\tilde{J}_\alpha(\nu A(t))}{A(t)^\alpha}\right| \quad \text{if } p \text{ is even}, \end{aligned} \tag{4.11}$$

where M_p and N_p are constants depending only on p. Recall that $\nu A(t)$ in our case is either positive or imaginary. Since α_p and β_p are continuous at $t = 0$, the estimates in (4.11) show that the error term ε_p has the same behavior as the first neglected term in the expansion (4.7) near both $t = 0$ and $t = -\infty$.

An alternative bound for the remainder ε_p could be expressed also in terms of the auxiliary functions $E_\alpha(x)$ and $M_\alpha(x)$ given in [13, Chap. 12, § 1.3] for real argument or $\mathfrak{E}_\alpha(z)$ and $\mathfrak{M}_\alpha(z)$ given in [13, Chap. 12, § 8] for complex argument.

The proof of (4.11) is divided into separate cases: (i) $0 \leqq t \leqq a < 1$, and (ii) $-\infty < t < 0$. We first consider case (i). Here, $0 \leqq A(t) < \pi/4$. We shall subdivide this case into two subcases: (i)(a) $0 \leqq \nu A(t) \leqq \delta$, and (i)(b) $\nu A(t) > \delta > 0$. In subcase (i)(a), we make the change of variable $\nu u = w$ in (4.8). The result is

$$\varepsilon_p = \left(-\frac{2}{\nu}\right)^p \frac{\nu^\alpha}{2\pi i} \int_{-\infty}^{(0+)} w^{-\alpha-1} h_p\left(\frac{w}{\nu}\right) \exp\left\{\frac{1}{2}\left(w - \frac{\nu^2 A^2}{w}\right)\right\} dw. \tag{4.12}$$

Since the last integral is bounded uniformly with respect to ν and t, (4.12) gives

$$\varepsilon_p = O(\nu^{-p+\alpha}), \tag{4.13}$$

the O-symbol being independent of t and ν. The estimate (4.11) now follows, when we take into account the small-z behavior of $\tilde{J}_{\alpha+1}(z)$. In subcase (i)(b), we recall that the curve EF in Fig. 3 is bounded away from the point D (see the second paragraph in § 3). Thus, we may take the radius ρ of the circle in the loop path of integration in (4.8) to be equal to $|A(t)|$. Consequently, we can write (4.8) as

$$\begin{aligned} \varepsilon_p = \left(-\frac{2}{\nu}\right)^p A(t)^{-\alpha} \Bigg\{ &\frac{\sin \alpha\pi}{\pi} \int_1^\infty r^{-\alpha-1} h_p(-A(t)r) \exp\left[\frac{\nu A(t)}{2}\left(\frac{1}{r} - r\right)\right] dr \\ &- \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\nu A(t) \sin\theta) h_p(A(t)\, e^{i\theta})\, e^{-i\alpha\theta}\, d\theta \Bigg\}. \end{aligned} \tag{4.14}$$

By Laplace's method [13, Thm. 7.1, p. 81], the first integral is asymptotically equal to $h_p(-A(t))/\nu A(t)$; and by the method of stationary phase [13, Thm. 13.1, p. 101], the second integral is asymptotic to

$$-2i\sqrt{\frac{2\pi}{\nu A}}\, h_p(iA(t)) \sin\left(\nu A - \frac{\pi}{4} - \frac{\alpha\pi}{2}\right).$$

Therefore we conclude

$$\varepsilon_p = \left(\frac{2}{\nu}\right)^p A(t)^{-\alpha} h_p(iA(t)) O\left(\frac{1}{\sqrt{\nu A(t)}}\right). \tag{4.15}$$

(Note that the contribution from the first integral in (4.14) actually cancels with the endpoint contribution from the second integral there.) The result (4.11) now follows from (4.15), in view of the large-z behavior of $\tilde{J}_\alpha(z)$.

We next consider case (ii). Here $A(t)$ is purely imaginary and $iA(t)$ is negative. We shall also divide this case into two subcases: (ii)(a) $-\eta \leqq t < 0$, and (ii)(b) $-\infty < t < -\eta$, where $\eta > 0$ is chosen so that $|A(t)| < \pi/2$ for $-\eta \leqq t < 0$ (cf. (4.10)). Since the argument for case (ii)(a) is similar to that of case (i), it will be omitted. However, we point out that in this case $\nu A(t)$ is purely imaginary and positive, and hence the roles of Laplace's method and the method of stationary phase in case (i) must be reversed. In case (ii)(b), we choose the radius ρ of the circle in the loop integral (4.8) to be any fixed positive number less than $\pi/2$ and make the substitution $u = -iA(t)v$. The result is

(4.16)

$$\varepsilon_p = \left(-\frac{2}{\nu}\right)^p (-iA(t))^{-\alpha} \frac{1}{2\pi i} \int_{-\infty}^{(0+)} v^{-\alpha-1} h_p(-iA(t)v) \exp\left\{-\frac{i\nu A(t)}{2}\left(v + \frac{1}{v}\right)\right\} dv.$$

Observe that in our present case, $iA(t)$ is large and negative, and that the saddle points are at $v = \pm 1$. By using Cauchy's theorem, we may make the circular portion of the contour in (4.16) have radius 1. An argument similar to that leading to (4.14) and (4.15), or a straightforward application of the saddle-point method [3, § 36], then gives

$$\varepsilon_p \sim \left(-\frac{2}{\nu}\right)^p [-iA(t)]^{-\alpha} \frac{1}{2\pi i} h_p(-iA(t)) \exp\{-i\nu A(t)\} \sqrt{\frac{-2\pi}{i\nu A(t)}}. \tag{4.17}$$

Note that $h_p(-iA(t))$ can be expressed in terms of the coefficients α_{2k} and β_{2k+1}, depending on whether p is even or odd. In view of the large-z behavior of $\tilde{J}_\alpha(z)$, z being purely imaginary and positive, the result in (4.11) now follows immediately from (4.17). This completes our proof of (4.11).

The leading coefficient α_0 in (4.7) can be calculated as follows. Since $\alpha_0 = h_0(iA) = h(iA)$, from (2.9) we have

$$\alpha_0 = \left[\frac{\sinh z(u)}{u}\right]^{-\alpha-1} \left.\frac{dz}{du}\right|_{u=iA}.$$

Since $u = iA$ corresponds to $z = z_+$ and $\sinh^2 z_+ = -t$, differentiating (1.6) twice with respect to u, we obtain

$$\left(\frac{dz}{du}\right)^2\Bigg|_{u=iA} = \frac{i \sinh^3 z_+}{tA \cosh z_+}.$$

Thus

$$\alpha_0 = \left(\frac{A}{\sqrt{t}}\right)^{\alpha+1}\left(\frac{t}{1-t}\right)^{1/4} A^{-1/2} \quad \text{if } 0 \leqq t < 1$$
$$= \left(\frac{|A|}{\sqrt{-t}}\right)^{\alpha+1}\left(\frac{-t}{1-t}\right)^{1/4} |A|^{-1/2} \quad \text{if } t < 0. \tag{4.18}$$

The second coefficient β_1 in (4.7) can also be computed, but the work is overwhelming. After a great deal of computation, we find that for $0 \leqq t < 1$

$$\beta_1 = \frac{\alpha_0 A}{2}\left[\frac{1-4\alpha^2}{8} A^{-1} + \frac{\sqrt{1-t}}{\sqrt{t}}\left\{\frac{4\alpha^2-1}{8} + \frac{1}{4}\frac{t}{1-t} + \frac{5}{24}\left(\frac{t}{1-t}\right)^2\right\}\right] \tag{4.19}$$

and for $t < 0$

$$\beta_1 = \frac{\alpha_0 A}{2}\left[\frac{1-4\alpha^2}{8} A^{-1} - i\frac{\sqrt{1-t}}{\sqrt{-t}}\left\{\frac{4\alpha^2-1}{8} + \frac{1}{4}\frac{t}{1-t} + \frac{5}{24}\left(\frac{t}{1-t}\right)^2\right\}\right], \tag{4.20}$$

where A is given by (2.7). Note that (4.20) can be formally obtained from (4.19) by simply replacing $\sqrt{t}$ by $i\sqrt{-t}$.

By using a matching procedure, it is possible to obtain recursive formulas for the coefficients α_n and β_n. This procedure consists of re-expanding the Bessel functions in (4.7) and comparing the results with expansions obtained by using the ordinary steepest descent method for t bounded away from zero. For some illustrations of this procedure, see [13, Chap. 11, Ex. 7.4 and Chap. 12, Ex. 5.2].

5. Airy function expansion. We now consider the case $b \leqq t < \infty$, where $0 < b < 1$. Our starting point is the integral representation [5, p. 190]

$$e^{-x/2} L_n^{(\alpha)}(x) = \frac{(-1)^n}{2^\alpha} \cdot \frac{1}{2\pi i} \int^{(1+)} e^{-xz/2}\left(\frac{1+z}{1-z}\right)^{\nu/4} (1-z^2)^{(\alpha-1)/2}\, dz, \tag{5.1}$$

where, as in (2.3), $\nu = 4n + 2\alpha + 2$. The path of integration encircles $z = 1$ in the positive direction, and closes at $\operatorname{Re} z = +\infty$, $|\operatorname{Im} z| = \text{constant}$. With $x = \nu t$, (5.1) can be written as

$$e^{-\nu t/2} L_n^{(\alpha)}(\nu t) = \frac{(-1)^n}{2^\alpha} \cdot \frac{1}{2\pi i} \int^{(1+)} e^{\nu f(z,\, t)} (1-z^2)^{(\alpha-1)/2}\, dz, \tag{5.2}$$

where

$$f(z, t) = \frac{1}{4} \ln \frac{1+z}{1-z} - \frac{1}{2} zt. \tag{5.3}$$

The saddle points of $f(z, t)$ are located at

$$z_+ = \sqrt{1-1/t} \quad \text{and} \quad z_- = -\sqrt{1-1/t}$$

if $t > 1$, and at

$$z_+ = i\sqrt{1/t-1} \quad \text{and} \quad z_- = -i\sqrt{1/t-1}$$

if $0 < t \leqq 1$. As $t \to 1$, the saddle points z_+ and z_- coalesce at $z = 0$. This suggests the use of the cubic transformation (1.2). Since $f(z, t)$ in (5.3) is an odd function, we set

$$\frac{1}{4} \ln \frac{1+z}{1-z} - \frac{1}{2} zt = \frac{u^3}{3} - B^2(t) u. \tag{5.4}$$

The coefficient $B(t)$ is to be chosen so that $z(u)$ is an analytic function of u. Upon differentiating (5.4) with respect to u and making the correspondence $u=+B(t)$ with $z=z_+$ and $u=-B(t)$ with $z=z_-$, we find

$$\begin{aligned} B(t) &= i[3\beta(t)/2]^{1/3}, \qquad 0<t\leqq 1, \\ &= [3\gamma(t)/2]^{1/3}, \qquad t>1, \end{aligned} \tag{5.5}$$

where

$$\begin{aligned} \beta(t) &= \tfrac{1}{2}[\cos^{-1} t^{1/2} - \sqrt{t-t^2}], \\ \gamma(t) &= \tfrac{1}{2}[(t^2-t)^{1/2} - \cosh^{-1} t^{1/2}]. \end{aligned} \tag{5.6}$$

Note that our function $B(t)$ and Erdélyi's $\phi(t)$ are related via the identity $B^2(t)=-\phi(t)$.

The mapping between z and u is most easily constructed by introducing an intermediate variable Z, defined by

$$\frac{1}{4}\ln\frac{1+z}{1-z} - \frac{1}{2}zt = Z = \frac{u^3}{3} - B^2(t)u. \tag{5.7}$$

We first consider the case $t>1$. The first quadrant of the z-plane is depicted in Fig. 4. Its image in the Z-plane is shown in Fig. 5. When z describes the indented boundary $ABCC'DEFA$ once, the image point Z also describes the corresponding boundary once. Here we again consider the line segments AB and BC being distinct parts of the boundary. Hence, the function $\xi(z)=\frac{1}{4}\ln((1+z)/(1-z))-\frac{1}{2}zt$ is one-to-one in that region (see [16, § 6.45]). The function $\eta(u)=\frac{1}{3}u^3-B^2(t)u$ is one-to-one in the shaded region $BCC'DEB$ in Fig. 6. This function has also been studied by Copson [3, p. 110–113]. The parametric equation of the boundary curve BE in Fig. 6 is given by

$$s^2-3r^2+3B^2=0 \qquad (s>0,\ r>0), \tag{5.8}$$

where $s=\operatorname{Im} u$ and $r=\operatorname{Re} u$. The transformation $z\leftrightarrow u$ is obtained by composing $\xi^{-1}: Z\to z$ and $\eta: u\to Z$. Since $z\leftrightarrow Z$ and $u\leftrightarrow Z$ are both one-to-one in the region $BCC'DEB$, so is $z\leftrightarrow u$. The mapping of the fourth quadrant in the z-plane can be discussed in the same manner, and the final result is obvious by symmetry.

The mappings $z\leftrightarrow Z$ and $Z\leftrightarrow u$, when $0<t<1$, can be constructed in a similar manner; they are illustrated in Figs. 4′, 5′, and 6′.

We now return to the integral (5.2). With the transformation (5.4) this integral becomes

$$e^{-\nu t/2}L_n^{(\alpha)}(\nu t) = \frac{(-1)^n}{2^\alpha}\cdot\frac{1}{2\pi i}\int_{\mathfrak{L}} h(u)\exp\left\{\nu\left(\frac{u^3}{3}-B^2(t)u\right)\right\}du, \tag{5.9}$$

where

$$h(u) = [1-z^2(u)]^{(\alpha-1)/2}\frac{dz}{du} \tag{5.10}$$

and $\mathfrak{L}$ is the branch of the hyperbolic curve in the right half-plane, half of which is given by (5.8). Note that (5.9) is first established separately under the conditions $t>1$ and $0<t<1$, and then extended to $t=0$ by continuity.

To derive the asymptotic expansion of the integral in (5.9), we apply the integration-by-parts technique introduced by Bleistein [1]. First, we write

$$h(u) = h_0(u) = \alpha_0 + \beta_0 u + (u^2 - B^2(t))g_0(u), \tag{5.11}$$

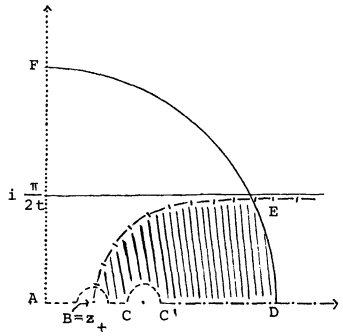


FIG. 4. *z-plane.*

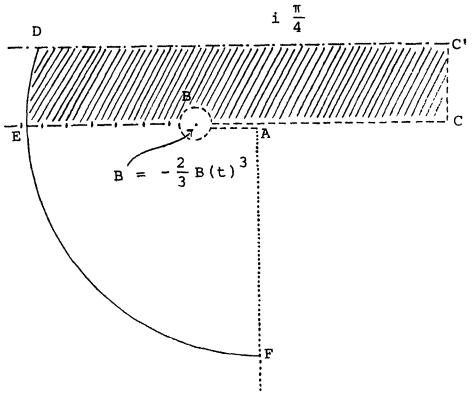


FIG. 5. *Z-plane.*

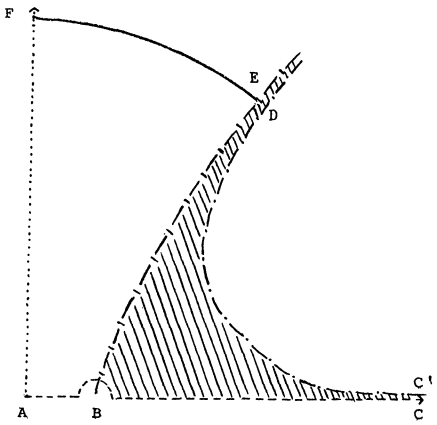


FIG. 6. *u-plane.*

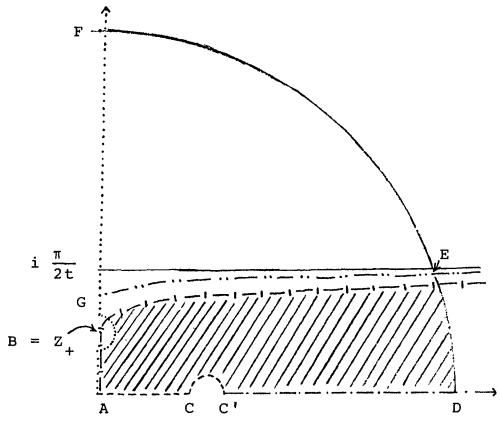


FIG. 4'. *z-plane* $(0<t<1)$.

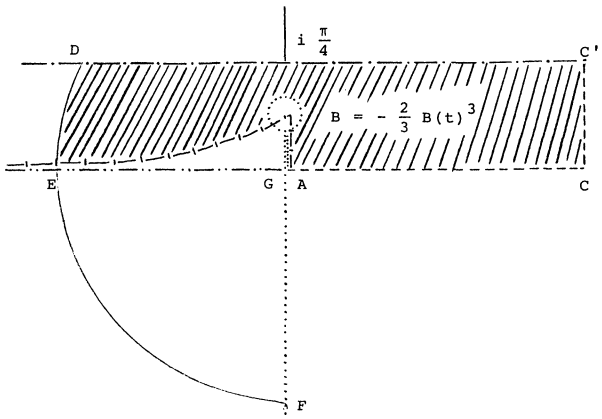


FIG. 5'. *Z-plane* $(0<t<1)$.

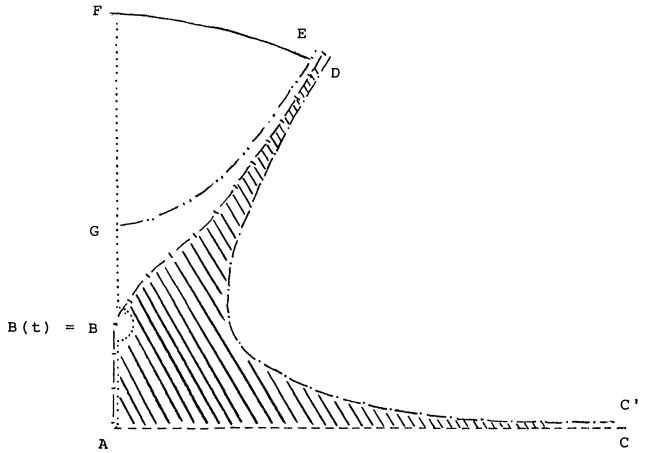


FIG. 6'. *u-plane* $(0<t<1)$.

where α_0, β_0, and $g_0(u)$ are to be determined. From (5.4), it is easily seen that $z(u)$ is an odd function of u. Hence, $h(u)$ is an even function of u, and $h(B)=h(-B)$. From this and (5.11), it follows that $\beta_0=0$, $\alpha_0=h_0(B)$, and

$$g_0(u)=\frac{h_0(u)-h_0(B)}{u^2-B^2}.$$

Thus, $g_0(u)$ is analytic in a neighborhood of $u=B$, and consequently analytic in the domain of $h_0(u)$. We now substitute (5.11) in (5.9), and express the first integral in terms of the Airy function. The second integral can be integrated by parts; the integrated term vanishes, since $g_0(u)$ has at most algebraic growth at infinity. The final result is

$$\begin{aligned}(-1)^n 2^\alpha e^{-\nu t/2} L_n^{(\alpha)}(\nu t) &= \mathrm{Ai}(\nu^{2/3}B^2)\frac{\alpha_0}{\nu^{1/3}} \\ &\quad -\left(\frac{1}{\nu}\right)\frac{1}{2\pi i}\int_{\mathfrak{L}} g_0'(u)\exp\left\{\nu\left(\frac{u^3}{3}-B^2u\right)\right\}du.\end{aligned} \tag{5.12}$$

The above procedure can be repeated, and proceeding in this manner we obtain

$$\begin{aligned}(-1)^n 2^\alpha e^{-\nu t/2} L_n^{(\alpha)}(\nu t) &= \mathrm{Ai}\,(\nu^{2/3}B^2)\sum_{k=0}^{[(p-1)/2]}\alpha_{2k}\nu^{-2k-1/3} \\ &\quad -\mathrm{Ai}'\,(\nu^{2/3}B^2)\sum_{k=0}^{[p/2]-1}\beta_{2k+1}\nu^{-2k-5/3}+\varepsilon_p,\end{aligned} \tag{5.13}$$

where we define inductively, for $m=0, 1, 2, \cdots$,

$$h_m(u)=\alpha_m+\beta_m u+(u^2-B^2)g_m(u), \qquad h_{m+1}(u)=g_m'(u). \tag{5.14}$$

It can be shown that for $n\geqq 0$, h_{2n} and g_{2n} are even analytic functions, and that h_{2n+1} and g_{2n+1} are odd analytic functions. Consequently,

$$\begin{aligned}&\alpha_{2n}=h_{2n}(B), \qquad \alpha_{2n+1}=0,\\ &\beta_{2n}=0, \qquad \beta_{2n+1}=h_{2n+1}(B)/B,\end{aligned} \tag{5.15}$$

and, for each n, both α_n and β_n are continuous in t for $0<t<\infty$. The remainder ε_p is explicitly given by

$$\varepsilon_p=\frac{1}{\nu^p}\cdot\frac{1}{2\pi i}\int_{\mathfrak{L}} h_p(u)\exp\left\{\nu\left(\frac{u^3}{3}-B^2u\right)\right\}du. \tag{5.16}$$

To estimate ε_p, we introduce the auxiliary functions

$$\tilde{\mathrm{Ai}}\,(z)=\begin{cases}\mathrm{Ai}\,(z) & \text{if } z\geqq 0,\\ [\mathrm{Ai}^2\,(z)+\mathrm{Bi}^2\,(z)]^{1/2} & \text{if } z<0,\end{cases} \tag{5.17}$$

and

$$\tilde{\mathrm{Ai}}'\,(z)=\begin{cases}\mathrm{Ai}'\,(z) & \text{if } z\geqq 0,\\ [\mathrm{Ai}'^2\,(z)+\mathrm{Bi}'^2\,(z)]^{1/2} & \text{if } z<0.\end{cases} \tag{5.18}$$

Furthermore, we define

$$\tilde{\alpha}_p(t)=\begin{cases}1 & \text{if } 0<t<\xi,\\ |\alpha_p| & \text{if } t>\xi,\end{cases}$$

and define $\tilde{\beta}_p(t)$ in a similar manner, where α_p and β_p are the coefficients in the expansion (5.13), and ξ is a positive number. The behavior of $\tilde{\mathrm{Ai}}\,(z)$ and $\tilde{\mathrm{Ai}}'\,(z)$ mimics

the behavior of $\mathrm{Ai}(z)$ and $\mathrm{Ai}'(z)$, respectively. Moreover, $\tilde{\alpha}_p$ and $\tilde{\beta}_p$ have the same behavior as α_p and β_p, respectively. For large values of ν and for $0<b\leqq t<\infty$, it can be shown that

$$|\varepsilon_p|\leqq\frac{N_p}{\nu^{p+2/3}}|\tilde{\beta}_p(t)||\tilde{\mathrm{A}}\mathrm{i}'(\nu^{2/3}B^2)| \quad \text{if } p \text{ is odd,} \tag{5.19}$$

and that

$$|\varepsilon_p|\leqq\frac{M_p}{\nu^{p+1/3}}|\tilde{\alpha}_p(t)||\tilde{\mathrm{A}}\mathrm{i}(\nu^{2/3}B^2)| \quad \text{if } p \text{ is even} \tag{5.20}$$

(for similar estimations, see [1] and [2]). Alternative estimates for ε_p could be expressed in terms of the auxiliary functions $E(x)$, $M(x)$, and $N(x)$ given in [13, Chap. 11, § 2] for a real argument or their analogues given in [13, Chap. 11, § 8] for a complex argument.

We conclude this paper with an evaluation of the leading coefficient α_0 in (5.13). From (5.15) and (5.10), it follows that

$$\alpha_0=[1-z^2(u)]^{(\alpha-1)/2}\left.\frac{dz}{du}\right|_{u=B}. \tag{5.21}$$

Since $u=B$ corresponds to $z=z_+$ and $1-z_+^2=1/t$, upon differentiating (5.4) twice, we obtain

$$\left.\left(\frac{dz}{du}\right)^2\right|_{u=B}=\frac{2B}{z_+t^2},$$

which in turn gives

$$\begin{aligned}\alpha_0&=t^{(1-\alpha)/2}\frac{\sqrt{2B}}{(t-1)^{1/4}t^{3/4}} \quad \text{if } t>1\\ &=t^{(1-\alpha)/2}\frac{\sqrt{2|B|}}{(1-t)^{1/4}t^{3/4}} \quad \text{if } 0<t<1.\end{aligned} \tag{5.22}$$

For completeness, we also record the formula

$$\beta_1=\frac{\alpha_0}{2B}\left[\frac{5}{24}B^{-3}-\frac{3}{4}\frac{\sqrt{t-1}}{\sqrt{t}}+\frac{1}{2}\frac{\sqrt{t}}{\sqrt{t-1}}-\frac{5}{12}\frac{t^{3/2}}{(t-1)^{3/2}}-(\alpha^2-1)\frac{\sqrt{t-1}}{\sqrt{t}}\right], \tag{5.23}$$

if $t>1$. A corresponding formula exists for the case $0<t<1$, and can be obtained formally from (5.23) by simply replacing $\sqrt{t-1}$ by $i\sqrt{1-t}$. To calculate the coefficients of higher terms, we can again use the matching procedure mentioned at the end of § 4.

6. Conclusion. In this paper two uniform asymptotic expansions are obtained for the Laguerre polynomial $L_n^{(\alpha)}(x)$. One is in terms of Bessel functions and holds for $-\infty<x\leqq a\nu$, while the other is in terms of Airy's integral and holds for $b\nu\leqq x<\infty$, where $0<b<a<1$ and $\nu=4n+2\alpha+2$. The leading terms of these expansions agree with the two asymptotic forms given by Erdélyi [4] using the theory of differential equations. Our method is based on two integral representations of the Laguerre polynomial.

From Olver's theory [13, Chaps. 11 and 12], it is evident that the two uniform asymptotic expansions for the Laguerre polynomial given in this paper can also be obtained from differential equation theory. However, there are functions which have relatively simple integral representations but do not satisfy any differential equations.

Thus it is desirable to have a method using integral representations as well as one using differential equation theory to derive uniform asymptotic expansions. The method presented in this paper is quite general, although we have considered only the case of Laguerre polynomials. We chose to do it this way, since in this case we could compare our results with those given by Erdélyi.

For both Bessel-function and Airy-function expansions, we have provided explicit expressions for the remainder terms. Despite this fact, only order estimates have been established for them. Our method does not seem to lend itself to the construction of numerical error bounds because of the complicated nature of the transformations involved. Therefore the integral-theoretic method is not as complete as the corresponding theory for differential equations [13, Chaps. 11 and 12], and error analysis for uniform asymptotic expansions of integrals obtained by the rational and the cubic transformations used in this paper remains a difficult problem yet to be resolved.

Acknowledgments. We are grateful to Professor F. W. J. Olver for making several helpful suggestions on an earlier version of the paper, and would also like to thank the referee for pointing out several shortcomings in the original version.

REFERENCES

[1] N. Bleistein, *Uniform asymptotic expansions of integrals with many nearly stationary points and algebraic singularities*, J. Math. Mech., 17 (1967), pp. 533-559.

[2] C. Chester, B. Friedman, and F. Ursell, *An extension of the method of steepest descents*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 599-611.

[3] E. T. Copson, *Asymptotic Expansions*, Cambridge University Press, London, 1965.

[4] A. Erdélyi, *Asymptotic forms for Laguerre polynomials*, J. Indian Math. Soc., Golden Jubilee Commemoration Volume, 24 (1960), pp. 235-250.

[5] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions*, Vol. 2, McGraw-Hill, New York, 1953.

[6] L. Hörmander, *An Introduction to Complex Analysis in Several Variables*, Van Nostrand, Princeton, NJ, 1966.

[7] R. E. Langer, *The asymptotic solutions of ordinary linear differential equations of the second order, with special reference to a turning point*, Trans. Amer. Math. Soc., 67 (1949), pp. 461-490.

[8] N. Levinson and R. M. Redheffer, *Complex Variables*, Holden-Day, San Francisco, 1970.

[9] A. I. Markushevich, *Theory of Functions of a Complex Variable*, 2nd ed., Chelsea, New York, 1977.

[10] B. Muckenhoupt, *Asymptotic forms for Laguerre polynomials*, Proc. Amer. Math. Soc., 24 (1970), pp. 288-292.

[11] F. W. J. Olver, *The asymptotic solution of linear differential equations of the second order for large values of a parameter*, Philos. Trans. Roy. Soc. London Ser. A, 247 (1954), pp. 307-327.

[12] ———, *The asymptotic solution of linear differential equations of the second order in a domain containing one transition point*, Philos. Trans. Roy. Soc. London Ser. A., 249 (1956), pp. 65-97.

[13] ———, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[14] G. Szegö, *Orthogonal Polynomials*, 3rd ed. Colloquium Publications, Vol. 23, American Mathematical Society, Providence, RI, 1967.

[15] N. M. Temme, *Special functions as approximants in uniform asymptotic expansions of integrals; a survey*, Rend. Sem. Mat. Univ. Politec. Torino, Fasc. spec. International Conference on Special Functions: Theory and Computation, 1985, pp. 289-317.

[16] E. C. Titchmarsh, *Theory of Functions*, 2nd ed., Oxford University Press, Oxford, 1939.

[17] F. Ursell, *Integrals with a large parameter: Paths and descent and conformal mapping*, Proc. Cambridge Philos. Soc., 67 (1970), pp. 371-381.

[18] M. Wyman, *An expansion of the Laguerre polynomials* $L_n^{\alpha}(z)$, Canad. Math. Bull., 5 (1962), pp. 229-240.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

020

AN ANALYTIC CONTINUATION FORMULA FOR THE GENERALIZED HYPERGEOMETRIC FUNCTION*

WOLFGANG BÜHRING†

Abstract. For $p=1,2,3,\cdots$, the hypergeometric function ${}_{p+1}F_p(z)$ is expressed in terms of power series in the variable $1/(z-z_0)$ that converge for $|z-z_0|>\max(|z_0|,|z_0-1|)$ where z_0 is any complex number.

Key words. special functions, hypergeometric functions, hypergeometric series, continuation formulas, analytic continuation

AMS(MOS) subject classifications. 33A30, 30B40

1. Introduction. Some of the analogues to the well-known continuation formulas of the Gaussian hypergeometric function ${}_2F_1(z)$ are considerably more complicated or are not known at all for the generalized hypergeometric functions ${}_{p+1}F_p(z)$ with $p=2,3,\cdots$. The coefficients in the expansion of the ${}_3F_2(z)$ in powers of $z-1$, for instance, although now known in detail [3], cannot be generalized in a simple way, but become more and more complicated as p increases. In contrast to this behavior, the expansion of ${}_3F_2(z)$ in powers of $1/(z-1)$, or more generally $1/(z-z_0)$ with any complex number z_0, is of such a relatively simple type that it can easily be written down for any p. This will be shown in the present paper.

2. The continuation formula. Our starting point is the continuation formula [4], [5]

$$\left(\prod_{j=1}^{p}\Gamma(b_j)\right)^{-1}{}_{p+1}F_p\left(\begin{matrix}a_1,a_2,\cdots,a_{p+1}\\ b_1,b_2,\cdots,b_p\end{matrix}\,\middle|\,z\right)=\sum_{k=1}^{p+1}C_k y_k(z), \tag{1}$$

with known connecting constants C_k to be displayed below in (9) and

$$y_k(z)=(-z)^{-a_k}{}_{p+2}F_{p+1}\left(\begin{matrix}1,a_k,1+a_k-b_1,1+a_k-b_2,\cdots,1+a_k-b_p\\ 1+a_k-a_1,1+a_k-a_2,\cdots,1+a_k-a_{p+1}\end{matrix}\,\middle|\,\frac{1}{z}\right), \tag{2}$$

$$|\arg(-z)|<\pi.$$

Here the ${}_{p+2}F_{p+1}$ is really a ${}_{p+1}F_p$ since one of its denominator parameters is always equal to 1. As they stand, the formulas are valid provided that none of the differences between any two numerator parameters a_j is equal to an integer.

In a similar way as in [2], we now may observe that, when $|z|$ is sufficiently large, the series representation of (2) may be re-expanded in powers of $1/(z-z_0)$ for any fixed z_0 by means of

$$\begin{aligned}(-z)^{-a_k-n}&=\left\{(z_0-z)\left(1+\frac{z_0}{z-z_0}\right)\right\}^{-a_k-n}\\ &=(z_0-z)^{-a_k-n}\sum_{j=0}^{\infty}\frac{(a_k+n)_j}{j!}\left(\frac{-z_0}{z-z_0}\right)^j.\end{aligned} \tag{3}$$

* Received by the editors May 18, 1987; accepted for publication August 17, 1987.
† Physikalisches Institut, Universität Heidelberg, D-6900 Heidelberg, Federal Republic of Germany.

Collecting the terms with equal powers of $1/(z-z_0)$ we then obtain

$$y_k(z) = (z_0 - z)^{-a_k} \sum_{n=0}^{\infty} D_n(a_k, z_0)(z - z_0)^{-n} \tag{4}$$

where

$$D_n(a_k, z_0) = \sum_{j=0}^{n} \frac{(1)_j (a_k)_j \prod_{m=1}^{p} (1 + a_k - b_m)_j}{j! \prod_{m=1}^{p+1} (1 + a_k - a_m)_j} \frac{(a_k + j)_{n-j}}{(n-j)!} (-z_0)^{n-j} \tag{5}$$

or, in view of

$$\frac{(a_k + j)_{n-j}}{(n-j)!} = (-1)^j \frac{(a_k)_n (-n)_j}{(a_k)_j n!}, \tag{6}$$

$$D_n(a_k, z_0) = \frac{(a_k)_n}{n!} (-z_0)^n {}_{p+2}F_{p+1}\left(\begin{matrix} 1, -n, 1 + a_k - b_1, 1 + a_k - b_2, \cdots, 1 + a_k - b_p \\ 1 + a_k - a_1, 1 + a_k - a_2, \cdots, 1 + a_k - a_{p+1} \end{matrix} \middle| \frac{1}{z_0} \right). \tag{7}$$

Here again the ${}_{p+2}F_{p+1}$ is really a ${}_{p+1}F_p$ since one of its denominator parameters is always equal to 1. Turning the (finite) series around [1], we may obtain a different representation of (7) as displayed below in (10). The convergence domain of the series (4) is determined by the finite singular points 0 or 1 of the differential equation of which the ${}_{p+1}F_p(z)$ and the $y_k(z)$ are solutions. Thus we have proved the following theorem.

THEOREM 1. *If no two of the numerator parameters a_j differ by an integer, we have for $|\arg(z_0 - z)| < \pi$ the continuation formula*

$$\left\{ \prod_{j=1}^{p} \Gamma(b_j) \right\}^{-1} {}_{p+1}F_p\left(\begin{matrix} a_1, a_2, \cdots, a_{p+1} \\ b_1, b_2, \cdots, b_p \end{matrix} \middle| z \right) = \sum_{k=1}^{p+1} C_k (z_0 - z)^{-a_k} \sum_{n=0}^{\infty} D_n(a_k, z_0)(z - z_0)^{-n} \tag{8}$$

where the connecting constants are

$$C_k = \frac{\prod_{j=1, j \neq k}^{p+1} \Gamma(a_j - a_k)}{\left(\prod_{j=1, j \neq k}^{p+1} \Gamma(a_j)\right)\left(\prod_{j=1}^{p} \Gamma(b_j - a_k)\right)} \tag{9}$$

and the power series with the coefficients

$$D_n(a_k, z_0) = \frac{(a_k)_n \prod_{j=1}^{p} (1 + a_k - b_j)_n}{\prod_{j=1}^{p+1} (1 + a_k - a_j)_n} {}_{p+1}F_p\left(\begin{matrix} a_1 - a_k - n, a_2 - a_k - n, \cdots, a_{p+1} - a_k - n \\ b_1 - a_k - n, b_2 - a_k - n, \cdots, b_p - a_k - n \end{matrix} \middle| z_0 \right) \tag{10}$$

converge outside the circle $|z - z_0| = \max(|z_0|, |z_0 - 1|)$.

Since one of the numerator parameters of the ${}_{p+1}F_p$ in (10) is always equal to $-n$, this ${}_{p+1}F_p$ is a polynomial in z_0 of degree n, so that the continuation formula (8) is relatively simple. With the choice $z_0 = 1$ it gives the desired generalization of one of the well-known continuation formulas of the Gaussian hypergeometric function ${}_2F_1(z)$.

Also of interest is the choice $z_0 = \frac{1}{2}$ since in this case the convergence domain, $|z - \frac{1}{2}| > \frac{1}{2}$, of the series on the right of (8) is significantly larger than it would be with $z_0 = 0$ or $z_0 = 1$ and contains in its interior the two exceptional points $z = \frac{1}{2}(1 \pm i\sqrt{3}) = \exp(\pm i\pi/3)$, which before were not accessible by power series. From this point of view, the present paper generalizes some results of [2].

REFERENCES

[1] W. N. BAILEY, *Generalized Hypergeometric Series*, Stechert-Hafner, New York, 1964.
[2] W. BÜHRING, *An analytic continuation of the hypergeometric series*, SIAM J. Math. Anal., 18 (1987), pp. 884–889.
[3] ———, *The behavior at unit argument of the hypergeometric function* ${}_3F_2$, SIAM J. Math. Anal., 18 (1987), pp. 1227–1234.
[4] Y. L. LUKE, *The Special Functions and Their Approximations, Vol.* 1, Academic Press, New York, 1969.
[5] N. NØRLUND, *Hypergeometric functions*, Acta Math., 94 (1955), pp. 289–349.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

021

ADDENDUM: Oscillation Theorems for Nonlinear Second-Order Differential Equations with a Nonlinear Damping Term*

S. R. GRACE†, B. S. LALLI‡, AND C. C. YEH§

This is an addendum to our previous paper [1]. For related results, we refer to [2], [3]. Consider the following second-order nonlinear differential equation

$$(E) \qquad (a(t)\psi(x(t))\dot{x}(t))^{\cdot} + q(t)f(x(t)) = 0 \qquad \left(\cdot = \frac{d}{dt}\right)$$

where

(a) $a \in C^1[0, \infty)$ and $a(t) > 0$ for $t \geqq 0$;

(b) $\psi \in C(-\infty, \infty)$, $\psi(x) > 0$ for all $x \in R$;

(c) $f \in C(-\infty, \infty) \cap C^1(-\infty, 0) \cap C^1(0, \infty)$, $xf(x) > 0$ for $x \neq 0$, and there exists a positive constant k such that $f'(x) \geqq k\psi(x) > 0$ for $x \neq 0$;

(d) $q \in C[0, \infty)$.

The following generalizes a theorem of Kusano, Onose, and Tobe [2].

THEOREM. *Suppose that there exists a function $\rho \in C^2([0, \infty), (0, \infty))$ with $\dot{\rho}(t) \geqq 0$ such that*

$$(1) \qquad \int^{\infty} \rho(t)q(t)\, dt = \infty,$$

$$(2) \qquad \int^{\infty} \frac{dt}{\rho(t)a(t)} = \infty,$$

$$(3) \qquad \int^{\infty} \frac{a(t)(\dot{\rho}(t))^2}{\rho(t)}\, dt < \infty,$$

$$(4) \qquad \int_{\varepsilon}^{\infty} \frac{\psi(u)}{f(u)}\, du < \infty \quad \text{and} \quad \int_{-\varepsilon}^{-\infty} \frac{\psi(u)}{f(u)}\, du < \infty \quad \text{for every } \varepsilon > 0.$$

Then all solutions of (E) *are oscillatory.*

Proof. Let $x(t)$ be a nonoscillatory solution of (E), say $x(t) > 0$ for $t \geqq t_1 \geqq 0$. Let

$$w(t) := \frac{a(t)\psi(x(t))\dot{x}(t)}{f(x(t))}\rho(t).$$

Then $w(t)$ satisfies

$$\dot{w}(t) = -\rho(t)q(t) + \frac{a(t)\dot{\rho}(t)\psi(x(t))\dot{x}(t)}{f(x(t))} - \frac{a(t)\rho(t)\psi(x(t))f'(x(t))(\dot{x}(t))^2}{(f(x(t)))^2}.$$

* Received by the editors August 3, 1987; accepted for publication December 18, 1987. SIAM J. Math. Anal., 15 (1984), pp. 1082–1093.

† Department of Mathematics, University of Petroleum and Minerals, Dhahran, Saudi Arabia.

‡ Department of Mathematics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0.

§ Department of Mathematics, Central University, Chung-Li, Taiwan, Republic of China.

Thus

$$
\begin{aligned}
w(t) \leqq w(t_1) &- \int_{t_1}^{t} \rho(s)q(s)\,ds + \int_{t_1}^{t} \frac{a(s)\dot{\rho}(s)\psi(x(s))\dot{x}(s)}{f(x(s))}\,ds \\
&- \int_{t_1}^{t} a(s)\rho(s)f'(x(s))\psi(x(s))\left(\frac{\dot{x}(s)}{f(x(s))}\right)^2 ds.
\end{aligned} \tag{5}
$$

It follows from the Schwarz inequality that

$$
\begin{aligned}
&\left|\int_{t_1}^{t} \frac{a(s)\dot{\rho}(s)\psi(x(s))\dot{x}(s)}{f(x(s))}\,ds\right| \\
&\quad \leqq \left(\int_{t_1}^{t} \frac{a(s)(\dot{\rho}(s))^2}{\rho(s)}\right)^{1/2}\left(\int_{t_1}^{t} \frac{a(s)\rho(s)\psi^2(x(s))(\dot{x}(s))^2}{(f(x(s)))^2}\,ds\right)^{1/2} \\
&\quad \leqq K\left(\int_{t_1}^{t} \frac{a(s)\rho(s)\psi^2(x(s))(\dot{x}(s))^2}{(f(x(s)))^2}\,ds\right)^{1/2},
\end{aligned}
$$

where $K := (\int_{t_1}^{\infty} (a(t)(\dot{\rho}(t))^2/\rho(t))\,dt)^{1/2}$ is finite because condition (3) holds. Thus, by (c), we have

$$
\begin{aligned}
w(t) \leqq w(t_1) &- \int_{t_1}^{t} \rho(s)q(s)\,ds + K\left(\int_{t_1}^{t} \frac{a(s)\rho(s)\psi^2(x(s))(\dot{x}(s))^2}{(f(x(s)))^2}\,ds\right)^{1/2} \\
&- k\int_{t_1}^{t} a(s)\rho(s)\psi^2(x(s))\left(\frac{\dot{x}(s)}{f(x(s))}\right)^2 ds.
\end{aligned}
$$

Clearly, the sum of the last two integrals on the right-hand side of the above inequality remains bounded above as $t \to \infty$. Thus, by (1), we have

$$
\lim_{t\to\infty} \frac{\rho(t)a(t)\psi(x(t))\dot{x}(t)}{f(x(t))} = -\infty.
$$

This means that there exists a $t_2 \geqq t_1$ such that

$$
\dot{x}(t) < 0 \quad \text{for } t \geqq t_2.
$$

This and (5) imply that there exists a $t_3 \geqq t_2$ such that

$$
1 + \int_{t_3}^{t} a(s)\rho(s)f'(x(s))\psi(x(s))\left(\frac{\dot{x}(s)}{f(x(s))}\right)^2 ds \leqq \frac{a(t)\rho(t)\psi(x(t))(-\dot{x}(t))}{f(x(t))}.
$$

The rest of the proof is similar to that of Theorem 7 in [1], and is omitted.

Remark. From condition (4), we see that our theorem and some theorems in [2] are strictly nonlinear oscillatory results for equation (E) with $\psi(x) = 1$, that is, these theorems cannot apply to equation $(a(t)\dot{x}(t))^{\cdot} + q(t)x(t) = 0$.

REFERENCES

[1] S. R. GRACE, B. S. LALLI, AND C. C. YEH, *Oscillation theorems for nonlinear second order differential equations with a nonlinear damping term*, SIAM J. Math. Anal., 15 (1984), pp. 1082–1093.

[2] T. KUSANO, H. ONOSE, AND H. TOBE, *On the oscillation of second order nonlinear ordinary differential equations*, Hiroshima Math. J., 4 (1974), pp. 491–499.

[3] J. S. W. WONG, *Oscillation thorems for second order nonlinear differential equations*, Bull. Inst. Math., Academia Sinica, 3 (1975), pp. 283–309.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

022

ERRATA: Homoclinic Orbits in Slowly Varying Oscillators*

STEPHEN WIGGINS† AND PHILIP HOLMES‡

Equation (3.14) is incorrect. The reason for this is that the perturbed orbits through the points $q^s_\varepsilon(0,\theta)$, $q^u_\varepsilon(0,\theta)\in\Pi_o$ cannot be assumed to lie on the same z level at $\varepsilon=0$. Rather they may lie on z levels differing by $O(\varepsilon)$. Thus it is possible for $z^s_1(\infty,\theta)\neq z^u_1(-\infty,\theta)$ with $q^s_\varepsilon(\infty,\theta)=q^u_\varepsilon(-\infty,\theta)$; for more discussion of this point, see Wiggins [1988]. Therefore (3.16) for the Melnikov function should be

$$M(\theta)=\int_{-\infty}^{\infty}(\nabla H\cdot g)(q_0(t),t+\theta)\,dt-\frac{\partial H}{\partial z}(\gamma(z_0))\int_{-\infty}^{\infty}g_3(q_0(t),t+\theta)\,dt. \tag{1}$$

In Proposition 5.1, (1) should be omitted and (2) should be changed to

$$\lim_{\substack{m\to\infty\\ z\to z_0}} M^{m/1}=\frac{1}{\|f(q_0(-\theta))\|}\left[\int_{-\infty}^{\infty}(\nabla H\cdot g)(q_0(t),t+\theta)\,dt-\frac{\partial H}{\partial z}(\gamma(z_0))\int_{-\infty}^{\infty}g_3(q_0(t),t+\theta)\,dt\right]. \tag{2}$$

Concerning example (6.1), the orbits (6.9a) and (6.9b) should be on the level $z=(-\gamma/\alpha)\sqrt{1-(\gamma/\alpha)}$ and the corrected Melnikov functions in (6.13) and (6.14) are, respectively, given by

$$M=-4\delta\left[\frac{d}{3}+\frac{\gamma b}{\sqrt{2}\alpha}\left(\frac{\pi}{2}\pm\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}\,b\right)\right]+2\gamma\left[2d-2\sqrt{2}b\left(\frac{\pi}{2}\pm\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}\,b\right)\right]\mp 2\sqrt{2}\pi\beta\frac{\sinh\left[\frac{1}{d}\sin^{-1}\sqrt{\frac{\alpha}{\gamma}}\,d\right]}{\sinh\frac{\pi}{d}}\cos\theta, \tag{3}$$

$$M=-4\delta\left[\frac{d}{3}+\frac{\gamma b}{\sqrt{2}\alpha}\left(\frac{\pi}{2}\pm\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}\,b\right)\right]+2\gamma\left[2d-2\sqrt{2}b\left(\frac{\pi}{2}\pm\sin^{-1}\sqrt{\frac{2\alpha}{\gamma}}\,b\right)\right]\pm 2\sqrt{2}\pi\beta\frac{\sinh\left[\frac{1}{d}\sin^{-1}\sqrt{\frac{\alpha}{\gamma}}\,d\right]}{\sinh\frac{\pi}{d}}\cos\theta. \tag{4}$$

Figure 6 is qualitatively correct; however, the details are not right and the corrected version can be found in Wiggins [1988].

* Received by the editors September 21, 1987; accepted for publication October 25, 1987. SIAM J. Math. Anal., 18 (1987), pp. 612-629.

† Department of Applied Mechanics, California Institute of Technology, Pasadena, California 91125.

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

Acknowledgments. We would like to acknowledge useful discussions with John Allen and Roger Samelson of Oregon State University and Kayo Ide of Caltech for recomputing the Melnikov functions.

REFERENCE

S. WIGGINS (1988), *Global Bifurcations and Chaos—Analytical Methods*, Springer-Verlag Applied Mathematical Sciences Series, Springer-Verlag, Berlin, New York, to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 5, September 1988

023

ERRATA: Restricted Quadratic Forms and Their Applications to Bifurcation and Stability in Constrained Variational Principles*

JOHN H. MADDOCKS†

The statement and proofs of Theorems 1 and 2 appearing in [1] are flawed. The blanket assumption that $\ker L$ is closed should be added to Hypotheses H1–H3, in which case $\mathscr{H} = \mathscr{R}(L) \oplus \ker L$, where $\mathscr{R}(L)$ denotes the range of L. Theorems 1 and 2 should then read:

THEOREM 1. *Let the subspaces $\mathscr{C}$, $L(\mathscr{C}^\perp)$, and $(\mathscr{C}^\perp \cap L^{-1}(\mathscr{C}))$ be closed, and let $\mathscr{C}^\perp \cap \mathscr{D}(L)$ be dense in $\mathscr{C}^\perp$. Then*

$$d^0(\mathscr{C}^\perp \cap \mathscr{D}) = \dim\{[\mathscr{C}^\perp \cap L^{-1}(\mathscr{C})] \backslash \ker L\}. \tag{2.3}$$

Remark. The hypotheses appear to be extremely restrictive, but they are valid in several important cases that arise in applications. In particular, $L(\mathscr{C}^\perp)$ is closed provided L is bounded away from zero on $(\ker L)^\perp = \mathscr{R}(L)$. Moreover, if L is bounded, $(\mathscr{C}^\perp \cap L^{-1}(\mathscr{C}))$ is closed whenever $\mathscr{C}$ is closed, and the condition involving $\mathscr{D}(L)$ is vacuous. When L is unbounded and only densely defined, the hypotheses are implied if $\mathscr{C}$ has finite dimension with $\mathscr{C} \subset \mathscr{D}(L)$.

Proof. Let $\mathscr{G}$ denote the subspace $(\mathscr{C}^\perp \cap L^{-1}(\mathscr{C})) \backslash \ker L$, and let $\mathfrak{M}$ be any maximal negative subspace of $\mathscr{C}^\perp \cap \mathscr{D}$. We shall prove that $\mathfrak{M} \oplus \mathscr{G}$ is a maximal nonpositive subspace of $\mathscr{C}^\perp \cap \mathscr{D}$. Equation (2.3) is a consequence of this fact because the sum of $\mathfrak{M}$ and $\mathfrak{G}$ is direct, and because

$$d^0(\mathscr{C}^\perp \cap \mathscr{D}) = \dim[\mathfrak{M} \oplus \mathscr{G}] - \dim \mathfrak{M}.$$

To see that the sum is direct note that any $x \in \mathscr{G}$ satisfies $x \in \mathscr{C}^\perp + \ker L$ and $Lx \in \mathscr{C}$. Consequently $Q(x)$ vanishes and therefore $x \notin \mathfrak{M}$. It is also clear that $\ker L \cap (\mathfrak{M} \oplus \mathscr{G}) = \{0\}$, and that $\mathfrak{M} \oplus \mathscr{G}$ is a nonpositive subspace.

It remains to prove that $\mathfrak{M} \oplus \mathscr{G}$ is maximal, that is, to demonstrate that

$$\text{If } x \in \mathscr{C}^\perp \cap \mathscr{D} \text{ is } L\text{-orthogonal to } \mathfrak{M} \oplus \mathscr{G}, \text{ and } x \notin \mathfrak{M} \oplus \mathscr{G} \oplus \ker L, \text{ then } \langle x, Lx \rangle > 0. \tag{2.4}$$

The maximality of $\mathfrak{M}$ as a negative subspace implies that any such x satisfies $\langle x, Lx \rangle \geqq 0$, so we obtain maximality of $\mathfrak{M} \oplus \mathscr{G}$ after reaching a contradiction on the assumption $\langle x, Lx \rangle = 0$. Because $(\mathscr{C}^\perp \cap L^{-1}(\mathscr{C}))$ is closed and is contained in $\mathscr{C}^\perp \cap \mathscr{D}$, any $x \in \mathscr{C}^\perp \cap \mathscr{D}$ can be decomposed as the sum

$$x = p + q, \quad \text{with } p \in \mathscr{C}^\perp \cap \mathscr{D} \cap [\mathscr{C}^\perp \cap L^{-1}(\mathscr{C})]^\perp \quad \text{and} \quad q \in [\mathscr{C}^\perp \cap L^{-1}(\mathscr{C})].$$

Moreover, $q \in \mathscr{G} \oplus \ker L$, so that

$$\text{if } x \notin \mathfrak{M} \oplus \mathscr{G} \oplus \ker L, \text{ then } p \notin \mathfrak{M} \oplus \mathscr{G} \oplus \ker L;$$

$$\text{if } x \text{ is } L\text{-orthogonal to } \mathfrak{M} \oplus \mathscr{G}, \text{ so is } p;$$

and, by choice of p and q,

$$\langle x, Lx \rangle = \langle p, Lp \rangle.$$

* Received by the editors January 6, 1988; accepted for publication January 11, 1988. SIAM J. Math. Anal., 16 (1985), pp. 47–68.

† Department of Mathematics, University of Maryland, College Park, Maryland 20742. The research of this author was supported by the Air Force Office of Scientific Research.

A contradiction on the assumption that there exists $p \neq 0 \in \mathscr{C}^{\perp} \cap \mathscr{D} \cap [\mathscr{C}^{\perp} \cap L^{-1}(\mathscr{C})]^{\perp}$ satisfying the hypotheses of (2.4) and $\langle p, Lp \rangle = 0$ is therefore sufficient to obtain the desired maximality.

The contradiction will be obtained from the construction of a vector that violates the maximality of the *negative* subspace $\mathfrak{M}$ of $\mathscr{C}^{\perp} \cap \mathscr{D}$. Because $\mathscr{C}^{\perp} \cap \mathscr{D}$ is dense in $\mathscr{C}^{\perp}$, a simple calculation shows that $L^{-1}(\mathscr{C}) = [L(\mathscr{C}^{\perp} \cap \mathscr{D})]^{\perp} \cap \mathscr{D}$. Consequently, because $\mathscr{C}$ and $L(\mathscr{C}^{\perp} \cap \mathscr{D})$ are closed,

$$(\mathscr{C}^{\perp} \cap L^{-1}(\mathscr{C}))^{\perp} = (\mathscr{C}^{\perp} \cap [L(\mathscr{C}^{\perp} \cap \mathscr{D})]^{\perp} \cap \mathscr{D})^{\perp} = (\mathscr{C}^{\perp} \cap [L(\mathscr{C}^{\perp} \cap \mathscr{D})]^{\perp})^{\perp}$$
$$= \mathscr{C} + L(\mathscr{C}^{\perp} \cap \mathscr{D}).$$

Thus, $p \in \mathscr{C}^{\perp} \cap \mathscr{D}$ can be represented as $p = u + Lv$ with $u \in \mathscr{C}$ and $v \in \mathscr{C}^{\perp} \cap \mathscr{D}$ with $Lv \neq 0$. Then a Gram-Schmidt procedure can be used to construct a vector $f = v - \sum_i \alpha_i u_i$ in $\mathscr{C}^{\perp} \cap \mathscr{D}$ that is L-orthogonal to $\mathfrak{M}$. Here $\{u_i, i = 1, \cdots, d^-(\mathscr{C}^{\perp} \cap \mathscr{D})\}$ is a mutually L-orthogonal basis of $\mathfrak{M}$ (the existence of which is guaranteed by Lemma 1), and $\alpha_i \in \mathfrak{R}$. Consequently, $\beta p + f \in \mathscr{C}^{\perp} \cap \mathscr{D}$ is L-orthogonal to $\mathfrak{M}$, for all $\beta \in \mathfrak{R}$. However, $Q(\beta p + f)$ can be expanded to obtain

$$\beta^2 \langle p, Lp \rangle + 2\beta \langle p, Lv \rangle - 2\beta \left\langle Lp, \sum_i \alpha_i u_i \right\rangle + \langle f, Lf \rangle. \tag{2.5}$$

By hypothesis, $\langle p, Lp \rangle = 0, \langle Lp, u_i \rangle = 0$, and, because $p \in \mathscr{C}^{\perp}$ with $u \in \mathscr{C}, \langle p, Ly \rangle = \langle p, u + Lv \rangle = \langle p, p \rangle$. Therefore (2.5) can be rewritten as $2\beta \langle p, p \rangle + \langle f, Lf \rangle$. Because $p \neq 0$, β can be chosen such that this last expression is negative, contradicting the maximality of $\mathfrak{M}$. □

The corrected statement of Theorem 2 is then:

THEOREM 2. *Under the hypotheses of Theorem* 1 *and the additional assumption that* $L^{-1}(\mathscr{C})$ *is closed,*

$$d^0(\mathscr{C}^{\perp} \cap \mathscr{D}) = \dim \{[\mathscr{C}^{\perp} \cap L^{-1}(\mathscr{C})] \backslash \ker L\} = d^0(L^{-1}(\mathscr{C})), \tag{2.6}$$

and

$$d^-(\mathscr{C}^{\perp} \cap \mathscr{D}) + d^-(L^{-1}(\mathscr{C})) + d^0(L^{-1}(\mathscr{C})) = \sigma^-. \tag{2.7}$$

Remark. If L is bounded, or $\mathscr{C}$ has finite dimension, $L^{-1}(\mathscr{C})$ is necessarily closed.

Proof. The proof requires only minor changes to that presented in [1]. First, the quantity $\cdots \cap (\ker L)^{\perp}$ should be replaced with $\cdots \backslash \ker L$ wherever it appears. Second, if $\mathscr{C} \not\subset \mathscr{R}(L)$, it is necessary to add elements of $\ker L$ to the vectors v constructed in (2.11) and (2.12) in order to guarantee that $v \in \mathscr{C}^{\perp}$. □

REFERENCE

[1] J. H. MADDOCKS, *Restricted quadratic forms and their application to bifurcation and stability in constrained variational principles*, SIAM J. Math. Anal., 16 (1985), pp. 47-68.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

ITERATES OF MAPS WITH SYMMETRY*

PASCAL CHOSSAT† AND MARTIN GOLUBITSKY‡

Abstract. In this paper the elementary aspects of bifurcation of fixed points, period doubling, and Hopf bifurcation for iterates of equivariant mappings are discussed. The most interesting of these is an algebraic formulation of the hypotheses of Ruelle's theorem (D. Ruelle [1973], "Bifurcations in the presence of a symmetry group," *Arch. Rational Mech. Anal.*, 51, pp. 136–152) on Hopf bifurcation in the presence of symmetry.

In the last sections this result is used to show that Hopf bifurcation from standing waves in a system of ordinary differential equations with $O(2)$ symmetry can lead directly to motion on an invariant 3-torus; indeed, depending on the exact symmetry of the standing waves, one might expect to see three invariant 3-tori emanating from such a bifurcation. The unexpected third frequency comes from drift along the torus of standing waves whose existence is forced by the $O(2)$ symmetry.

Key words. symmetry, Hopf bifurcation, iterates of mappings

AMS (MOS) subject classifications. 58F14, 58F27, 34C35

Introduction. Symmetries change the types of bifurcation that may be expected in discrete dynamical systems. Typically, nonsymmetric systems generate unique branches of new solutions at points of bifurcation while symmetric systems generate multiple branches. Results of Vanderbauwhede [1980] and Golubitsky and Stewart [1985] on steady-state and Hopf bifurcation in continuous systems show that certain of these solution branches may be enumerated using only group theoretic techniques. The first task in this paper is the translation of these results to statements about bifurcation in the discrete dynamics of equivariant mappings. For further background, see Field [1980], [1986].

In § 1, we briefly describe the group theoretic results of Vanderbauwhede [1980] and Golubitsky and Stewart [1985]. In § 2, we apply Vanderbauwhede's result in a straightforward manner to enumerate certain branches of fixed points and branches of period two orbits for equivariant mappings. We also indicate how the simplest nontrivial symmetry ($\mathbb{Z}_2 = \{\pm 1\}$ acting on $\mathbb{R}$) may be expected to affect period doubling cascades and lead naturally to mergings of attractors. In § 3, we adapt the results of Golubitsky and Stewart [1985] to enumerate branches of invariant curves stemming from Hopf bifurcation of equivariant mappings. This adaptation leans heavily on nontrivial results of Ruelle [1973]. Our contribution is really only to observe that there is an algebraic formulation for Hopf bifurcation of equivariant maps that satisfies the hypotheses of Ruelle's theorem.

The second task in this paper is to enumerate the number and type of tori that are produced when a periodic solution to an equivariant system of ordinary differential equations (ODEs) loses stability by having Floquet multipliers cross the unit circle in the complex plane. For example, we show in § 4 that (under certain hypotheses) standing wave solutions to $O(2)$ symmetric systems generate (generically) three branches of 3-tori at such a bifurcation. The existence of this extra frequency comes

* Received by the editors March 21, 1987; accepted for publication (in revised form) December 20, 1987.

† I.M.S.P., Université de Nice, Parc Valrose, F-06034 Nice Cedex, France. The research of this author was supported in part by the Applied Computational Mathematics Program of the Defense Advanced Research Projects Agency.

‡ Department of Mathematics, University of Houston, University Park, Houston, Texas 77004. The research of this author was supported in part by the Applied Computational Mathematics Program of the Defense Advanced Research Projects Agency, by National Aeronautics and Space Administration–Ames grant NAG2-279, and by National Science Foundation grant DMS-8402604.

from the $O(2)$ symmetries and is based on observations of Iooss [1986] and Chossat [1986]. In § 5, we give a general setting for the example in § 4.

1. Background. Let $\Gamma \subset O(n)$ be a compact Lie group acting linearly on $\mathbb{R}^n$ and let $f:\mathbb{R}^n \to \mathbb{R}^n$ be a one-parameter family of smooth mapping commuting with Γ, i.e.,

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

The equivariant branching lemma of Vanderbauwhede [1980] and Cicogna [1981] gives a simple algebraic condition for determining the existence of branches of steady-state solutions to the system of ODEs

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

We assume that Γ acts absolutely irreducibly on $\mathbb{R}^n$, that is, that the only linear maps on $\mathbb{R}^n$ that commute with Γ are scalar multiples of the identity. For a subgroup Σ, we define

$$\text{Fix}(\Sigma) = \{y \in \mathbb{R}^n : \sigma y = y \quad \forall \sigma \in \Sigma\}. \tag{1.2}$$

Applying the chain rule to (1.1) implies that

$$(df)_{0,\lambda}\gamma = \gamma(df)_{0,\lambda}.$$

Absolute irreducibility then implies that

$$(df)_{0,\lambda} = c(\lambda)I. \tag{1.3}$$

Also observe that

$$f: \text{Fix}(\Sigma) \times \mathbb{R} \to \text{Fix}(\Sigma) \tag{1.4}$$

since $\sigma f(y, \lambda) = f(\sigma y, \lambda) = f(y, \lambda)$ for all $\sigma \in \Sigma$, $y \in \text{Fix}(\Sigma)$. In particular, irreducibility implies $\text{Fix}(\Gamma) = \{0\}$, and hence by (1.4)

$$f(0, \lambda) = 0.$$

Thus, there is a "trivial" solution at $x = 0$.

EQUIVARIANT BRANCHING LEMMA. *Let $\Sigma \subset \Gamma$ be a subgroup. Assume that $c(0) = 0$, $c'(0) \neq 0$, and $\dim \text{Fix}(\Sigma) = 1$. Then there exists a unique (nontrivial) branch of small amplitude steady states for $f|(\text{Fix}(\Sigma) \times \mathbb{R}) = 0$.*

See Ihrig and Golubitsky [1984], Golubitsky, Swift, and Knobloch [1984], and Golubitsky, Stewart, and Schaeffer [1988] for applications of this result.

There is a similar result regarding Hopf bifurcation in symmetric systems. Here we assume that the system $\dot{x} = f(x, \lambda)$ is on the center manifold. In particular, we assume that $x \in \mathbb{R}^{2n}$ and that

$$L \equiv (df)_{0,0} = \begin{pmatrix} 0 & -\omega I_n \\ \omega I_n & 0 \end{pmatrix}.$$

There is the natural action of the circle group S^1 or $\mathbb{R}^{2n}$ given by

$$x \to \exp(tL)x. \tag{1.5}$$

We assume that the action of $\Gamma \times S^1$ on $\mathbb{R}^{2n}$ is irreducible. It then follows, as above, that $f(0, \lambda) \equiv 0$, i.e., that there is a "trivial" steady-state solution. It also follows that the eigenvalues of $(df)_{0,\lambda}$ are $\sigma(\lambda) \pm i\omega(\lambda)$, each of multiplicity n.

THEOREM 1.1. *Let* $\Sigma \subset \Gamma \times S^1$ *be a subgroup. Assume that* $\sigma(0)=0$, $\omega(0)\neq 0$, $\sigma'(0)\neq 0$, *and* $\dim \operatorname{Fix}(\Sigma)=2$. *Then there exists a unique (nontrivial) branch of small amplitude periodic trajectories with period near* $2\pi/\omega(0)$ *to* $\dot{x}=f(x,\lambda)$ *with symmetries* Σ.

Note $(\sigma,\theta)\in\Sigma\subset\Gamma\times S^1$ is a symmetry of a periodic solution $x(t)$ to $\dot{x}=f(x,\lambda)$ if

$$\gamma x(t)=x(t+\theta). \tag{1.6}$$

See Golubitsky and Stewart [1985], [1986b], Roberts, Swift, and Wagner [1986], and Golubitsky, Stewart, and Schaeffer [1988] for a proof and applications.

We remark that in certain instances it is possible to use invariant theory and group theory to compute the asymptotic stability of the steady-state and periodic solutions found using the results stated above. We refer to these references for examples of this process.

2. Fixed points and period doubling. Let $g:\mathbb{R}^n\times\mathbb{R}\to\mathbb{R}^n$ be a one-parameter family of Γ-equivariant mappings. We assume that Γ acts absolutely irreducibly on $\mathbb{R}^n$. It follows that $x=0$ is a "trivial" fixed point for g and that $(Dg)_{0,\lambda}=c(\lambda)I$. In this section, we briefly discuss the bifurcation of fixed points ($c(0)=+1$) and period doubling bifurcation ($c(0)=-1$).

LEMMA 2.1. *Let* $\Sigma\subset\Gamma$ *be a subgroup. Suppose that* $c(0)=1$, $c'(0)\neq 0$, *and* $\dim \operatorname{Fix}(\Sigma)=1$. *Then* $g(x,\lambda)$ *has a unique (nontrivial) branch of fixed points in* $\operatorname{Fix}(\Sigma)$.

Proof. Set $f(x,\lambda)=g(x,\lambda)-x$ and apply the Equivariant Branching Lemma. □

To eliminate trivialities, we assume that Σ is an isotropy subgroup, that is, there is an $x\in\mathbb{R}^n$ such that

$$\Sigma=\{\gamma\in\Gamma:\ \gamma x=x\}. \tag{2.1}$$

The largest subgroup of Γ that leaves $\operatorname{Fix}(\Sigma)$ invariant is $N(\Sigma)$, the normalizer of Σ in Γ (cf. Golubitsky [1983] or Golubitsky, Stewart, and Schaeffer [1988]). It follows that $g|\operatorname{Fix}(\Sigma)\times\mathbb{R}$ commutes with the action of $N(\Sigma)/\Sigma$. Now, if we assume that Σ is a maximal isotropy subgroup (a hypothesis that is satisfied when $\dim \operatorname{Fix}(\Sigma)=1$), then $N(\Sigma)/\Sigma$ acts fixed point free. It follows that when $\dim \operatorname{Fix}(\Sigma)=1$, either $N(\Sigma)=\Sigma$ or $N(\Sigma)/\Sigma\cong\mathbb{Z}_2$. In the latter case, the bifurcation of fixed points in Lemma 2.1 will be via a pitchfork bifurcation, with the two new bifurcating fixed points lying on the same group orbit (conjugacy being given by any $\gamma\in N(\Sigma)\sim\Sigma$). When $N(\Sigma)=\Sigma$, we expect each new fixed point to be on a distinct group orbit.

We now discuss the case of period doubling, i.e., $c(0)=-1$. As in the standard period doubling theorem (without symmetry), we observe that nonzero fixed points for the composite mapping g^2 correspond to period two points for g, since the implicit function theorem guarantees that there are no new fixed points for g. We apply Lemma 2.1 to g^2 to obtain the following lemma.

LEMMA 2.2. *Let* $\Sigma\subset\Gamma$ *be a subgroup. Suppose that* $c(0)=-1$, $c'(0)\neq 0$, *and* $\dim \operatorname{Fix}(\Sigma)=1$. *Then* $g(x,\lambda)$ *has a unique branch of period two points in* $\operatorname{Fix}(\Sigma)$.

Again, we have different interpretations for Lemma 2.2, depending on whether $N(\Sigma)=\Sigma$ or $N(\Sigma)/\Sigma\cong\mathbb{Z}_2$. In the first case, we expect a standard period doubling to occur, while in the second case, the equivariance of $g|(\operatorname{Fix}(\Sigma)\times\mathbb{R})$ with respect to $\mathbb{Z}_2$ implies that

$$g(x,\lambda)=-x. \tag{2.2}$$

To verify (2.2), note that $g(-x,\lambda)=-g(x,\lambda)$. Therefore, if x is a period two point for

g, then so is $-x$. Since x and $-x$ are in Fix (Σ) and the period two orbit obtained from Lemma 2.2 in Fix (Σ) is unique, it follows that $g(x, \lambda) = -x$.

Remark. Identity (2.2) states that this period two trajectory is a discrete analogue of a rotating wave; the same result is obtained by taking one timestep (iteration by g) or by acting by the group $(x \to -x)$.

We end this section with some speculations on period doubling sequences when $N(\Sigma)/\Sigma \cong \mathbb{Z}_2$. As a parameter is varied, we might expect the trivial fixed point to undergo a bifurcation to a nontrivial fixed point, as in Lemma 2.1. As we discussed above, when $N(\Sigma)/\Sigma \cong \mathbb{Z}_2$, this new fixed point is formed by a pitchfork bifurcation.

Suppose now that, as this parameter is varied, each of the nontrivial fixed points undergoes a period doubling sequence. The $\mathbb{Z}_2$ action forces the period doubling sequence to occur at the same parameter values for each nontrivial fixed point. The simplest such example is given by the cubic polynomial

$$g(x, \lambda) = \mu x - x^3, \qquad \mu > 0 \tag{2.3}$$

on Fix $(\Sigma) \times \mathbb{R}$. For $\mu > 0$, each of these period doubling sequences seems to behave like the simple logistic equation. This results in pairs of attractors (one for $x > 0$ and one for $x < 0$) consisting of single orbits filling up parts of the real line, say, for x in $[\alpha, \beta]$ and for x in $[-\beta, -\alpha]$.

As λ is increased, α decreases and eventually becomes negative (when $\lambda = 3\sqrt{3/2}$). This merging of attractors causes an interesting kind of chaotic behavior. Start with an initial point $x_0 > 0$ and form the iterated sequence $x_{n+1} = g(x_n, \lambda)$. Now form the symbol sequence of $+$'s and $-$'s where the nth element in the sequence is sgn (x_n). In effect, we see chaotic behavior on two time scales. There is the chaotic behavior on a fast time scale within each of the attractors ($[0, \beta]$ and $[-\beta, 0]$) and then there is the chaotic behavior on a slow time scale defined by the transitions between the remnants of the two attractors.

A detailed study of the related map

$$h(x, \lambda) = -(\mu x - x^3), \qquad \mu > 0$$

is given in Rogers and Whitley [1983]. There, however, the primary bifurcation of the fixed point $x = 0$ is a period doubling bifurcation, as discussed in Lemma 2.2.

3. Hopf bifurcation. In this section, we assume that the trivial fixed point for the Γ-equivariant mapping g loses stability by a pair of complex conjugate eigenvalues crossing the unit circle. Due to the presence of symmetry, the eigenvalues may have high multiplicity. We assume that $g: \mathbb{R}^{2n} \times \mathbb{R} \to \mathbb{R}^{2n}$ and that $(Dg)_{0,0}$ has eigenvalues $e^{\pm 2\pi i\theta}$, each with multiplicity n, where $\theta \neq 0, \frac{1}{2}$.

The standard Hopf bifurcation theorem for mappings ($n = 1$) states that if $\theta \neq \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$ and if the eigenvalues cross the unit circle with nonzero speed, then there exists a family of invariant circles for $g(\cdot, \lambda)$ emanating from the trivial fixed point $x = 0$. This theorem is proved using near identity changes of coordinates to put the terms of g up to order four in normal form. This truncated normal form actually has S^1 symmetry, and, because of this symmetry, we can easily find invariant circles using polar coordinates. Then we use scaling and normal hyperbolicity arguments to show that the invariant circles that are present at order four persist independently of the higher order terms in g. When resonances exist ($\theta = \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$), the normal form does not have this S^1-equivariance in the fourth-order truncated normal form (cf. Arnold [1977], [1983] and Iooss [1979]).

We obtain a simple generalization of the Hopf bifurcation theorem as follows: Let $\Sigma \subset \Gamma$ be a subgroup with dim Fix $(\Sigma) = 2$. Then there exists a branch of g-invariant

circles in Fix (Σ), as long as $\theta \neq \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$. Just apply the standard Hopf theorem to $g|\text{Fix}(\Sigma)$; the eigenvalues of $D(g|\text{Fix}(\Sigma))$ are constrained by dimension to be simple.

Remark. Assume that Σ is an isotropy subgroup with dim Fix $(\Sigma) = 2$. The group $N(\Sigma)/\Sigma$ acts on Fix (Σ) by a fixed-point free action and $g|\text{Fix}(\Sigma)$ commutes with this action. (In fact, Σ is a maximal isotropy subgroup since the complex eigenvalues preclude the existence of one-dimensional fixed-point subspaces; cf. Golubitsky and Stewart [1985].) Fixed-point free actions on $\mathbb{R}^2$ exist only for the groups 1, $\mathbb{Z}_n$ $(n \geqq 2)$ or $SO(2)$. Observe that if $N(\Sigma)/\Sigma \cong \mathbb{Z}_n$ $(n \geqq 5)$ or $SO(2)$, then $g|\text{Fix}(\Sigma)$ automatically has a fourth-order truncated normal form with S^1 symmetry. In these cases, the assumption $\theta \neq \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$ is *not* necessary, as the remainder of the proof of the standard Hopf theorem is still valid.

As in the case of Hopf bifurcation for systems of ODEs (Theorem 1.1), we can improve on this simple generalization by looking for subgroups of $\Gamma \times S^1$ with two-dimensional fixed-point subspaces. First, we define the action of S^1. Choose a matrix A with purely imaginary eigenvalues such that $e^A = (dg)_{0,0}$. The action of S^1 is then given by e^{tA}. Since $(dg)_{0,0}$ commutes with Γ, so does the action of S^1. In this way, we have defined an action of $\Gamma \times S^1$ on $\mathbb{R}^{2n}$.

THEOREM 3.1. *Let* $\Sigma \subset \Gamma \times S^1$ *be a subgroup such that* dim Fix $(\Sigma) = 2$. *Assume* $\theta \neq \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$ *and that the eigenvalues cross the unit circle with nonzero speed. Then generically there exists a unique branch of g-invariant circles emanating from the trivial fixed point* $x = 0$ *and this branch is tangent to* Fix $(\Sigma) \subset \mathbb{R}^{2n} \times \mathbb{R}$ *at* $x = 0$.

Proof. The truncated normal form h of g has symmetry group $\Gamma \times S^1$. Therefore, $h: \text{Fix}(\Sigma) \times \mathbb{R} \to \text{Fix}(\Sigma)$, and we can find invariant circles for h, as above. At this point, however, we cannot conclude directly from the proof of the standard Hopf bifurcation theorem that there is a family of g-invariant circles. The difficulty is that g itself need not leave invariant Fix (Σ) since g does not necessarily commute with S^1. However, Ruelle [1973, Thm. 3.1] does prove a theorem sufficient to conclude that g has a family of invariant circles, at least when certain assumptions, which are valid generically, hold.

The needed assumptions are the following:

(a) The third-order terms in h determine the direction of branching of the invariant circles of h in Fix (Σ).

(b) The invariant circles for h are normally hyperbolic in the sense that the eigenvalues of dh on the invariant circles, which are not forced by the $\Gamma \times S^1$ action to be unity, in fact lie off the unit circle.

In order for (b) to hold, it is often necessary to have truncated the normal form at some high order. This order depends on both $\Gamma \times S^1$ and the subgroup Σ. Once the invariant circles of h are normally hyperbolic, Ruelle's Theorem 3.1 is sufficient to prove that the higher order terms of g (which are not in normal form) will neither destroy the invariant circles nor change their stability. □

Example 3.2. Consider $\Gamma = D_n$ $(n \geqq 3)$ acting absolutely irreducibly on $\mathbb{C}$ and by the diagonal action on $\mathbb{C}^2$. As was shown in Golubitsky and Stewart [1986b], there are three (conjugacy classes of) isotropy subgroups in $D_n \times S^1$ where the fixed-point subspaces are two-dimensional. Theorem 3.1 implies that for D_n-equivariant mappings, we may expect three families of g-invariant circles at such a Hopf bifurcation. We note that two of the isotropy subgroups are isomorphic to $\mathbb{Z}_2$ and one to $\mathbb{Z}_3$. The normal hyperbolicity of the $\mathbb{Z}_3$ circles are determined at third order, while the normal hyperbolicity of the $\mathbb{Z}_2$ branches are determined at order m where

$$m = \begin{cases} n, & n \text{ odd}, \\ (n+2)/2, & n \text{ even}. \end{cases}$$

4. Bifurcation of standing waves to 3-tori. For an $O(2)$ invariant system of ODEs, a symmetry-breaking Hopf bifurcation leads to two types of periodic solutions: rotating waves and standing waves (cf. Golubitsky and Stewart [1985]). We are interested here in the bifurcation of these periodic solutions to tori. Bifurcation from rotating waves has been considered by Rand [1982], Renardy [1982], and Iooss [1984]. By changing coordinates to a rotating frame, they show that rotating waves correspond to stationary solutions and that 2-tori may be found by standard Hopf bifurcation techniques for systems of ODEs. Moreover, the circular symmetry of the rotating waves forces the flow on the 2-torus to be linear. Standing waves, however, have no circular symmetry in their isotropy subgroup, and the technique of changing coordinates to a rotating frame does not apply. Using the techniques described in § 3 applied to a certain Poincaré map, we shall study here the bifurcation to tori from standing waves. In the next section, we give a unified discussion of these two techniques when $O(2)$ is replaced by a general group Γ.

Bifurcation to 2-tori from a branch of standing waves has been considered in the context of degenerate, symmetry-breaking, $O(2)$ Hopf bifurcations by a number of authors (Erneux and Matkowsky [1984], Knobloch [1986], and Golubitsky and Roberts [1987]). These authors decouple the normal form equations for $O(2)$ Hopf bifurcation (on $\mathbb{C}^2$) into phase-amplitude equations and find the 2-tori by steady-state bifurcation in the amplitude equations. Using the extra S^1 phase shift symmetry of normal form, it is easy to see that in normal form the flow on these 2-tori must also be linear. Chossat [1986] uses a Lyapunov–Schmidt reduction to prove that the flow on such 2-tori is linear even when the vector field is not assumed to be in normal form. His method is to assume that the flow on the 2-torus has the form $y(t) = R_{\eta t}x(t)$ where x is periodic, η is a real parameter, and R_θ denotes the action of θ in $SO(2)$. The original equation is then transformed by substitution of $y(t)$ and elimination of $R_{\eta t}$ to an equation for x. It is this equation to which the Lyapunov–Schmidt reduction is applied, and this idea we will use to analyze bifurcation to tori from standing waves. In § 5, we will show that, in principle, when considering bifurcation to tori from a branch of periodic solutions in a symmetric system, one of the two techniques described above always works. Which one will work depends on the symmetries of the periodic solution.

Consider the following system of ODEs:

$$\dot{y} = F(y, \lambda), \qquad F(0, \lambda) = 0 \tag{4.1}$$

where $y \in \mathbb{R}^N$ and $F: \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ commutes with a linear action of $O(2)$ on $\mathbb{R}^N$. This action may not be faithful; we assume, however, that the kernel of the action is the cyclic group $\mathbb{Z}_n$, $n \geqq 1$.

Let $y(t)$ be a *standing wave* periodic solution to (4.1), that is, assume that the isotropy subgroup

$$\Sigma = \{\gamma \in O(2): \gamma y(t) = y(t)\} \tag{4.2}$$

is discrete and contains a reflection in $O(2)$. Thus $\Sigma = D_n$. Note that standing waves lie on the invariant 2-torus

$$M = \{\gamma y(t): \gamma \in O(2)\}$$

foliated by periodic trajectories.

We now consider bifurcation of standing waves to tori. This bifurcation is detected by having a complex conjugate pair of Floquet multipliers cross the unit disk at $e^{\pm 2\pi i\theta}$ where $\theta \neq 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$. The eigenspaces corresponding to these Floquet multipliers are invariant under $\Sigma = D_n$, and generically we may assume that D_n acts irreducibly

on the eigenspaces. The irreducible representations of D_n are either one-dimensional or, if $n \geqq 3$, two-dimensional.

We prove the following theorem.

THEOREM 4.1. *Let $x_\lambda(t)$ be a family of standing wave periodic solutions to the $O(2)$ symmetric system* (4.1) *with isotropy subgroup D_n. Assume that the periodic solution loses stability by having a pair of complex conjugate Floquet multipliers cross the unit disk with nonzero speed and assume that D_n acts irreducibly on the corresponding eigenspaces.*

(a) *If the Floquet multipliers are simple, then there exists a branch of 3-tori emanating from this bifurcation.*

(b) *If the Floquet multipliers are double (which may happen generically when $n \geqq 3$), then there exist three branches of 3-tori emanating from the bifurcation.*

Our proof consists of constructing a D_n-equivariant Poincaré map to which we can apply the results of § 3.

Remarks. (a) Such bifurcations to 3-tori occur in the interaction of two symmetry-breaking $O(2)$ Hopf bifurcations (see Chossat, Golubitsky, and Keyfitz [1986]) and in the interaction of $O(2)$ symmetry-breaking steady-state and Hopf bifurcations (see Golubitsky and Stewart [1986a]).

(b) Normally we would expect the bifurcation of a periodic solution to tori to produce an invariant 2-torus. The extra frequency comes from the $O(2)$ symmetry. As noted above, each standing wave $x(t)$ lies on the 2-torus M defined by $\gamma x(t)$ for $\gamma \in O(2)$. When bifurcation to tori occurs, we get two independent frequencies from the "Poincaré map" and a third independent (slow) frequency from flow transverse to $\gamma x(t)$ in the group generated 2-torus M. It is here that we use the ideas of Iooss [1986] and Chossat [1986], described above.

(c) Suppose that the standing waves are generated by Hopf bifurcation with $O(2)$ symmetry from an invariant steady state in (4.1). Then the bifurcation to tori we describe in Theorem 4.1 cannot occur in a system of differential equations posed only on the four-dimensional center subspace. Since the hypotheses of the theorem presume the existence of four nontrivial Floquet multipliers and periodic solutions always have one trivial Floquet multiplier (equal to unity), such a system cannot live on a four-dimensional manifold. In effect, the question we discuss here is: suppose that a standing wave with D_n symmetry is formed from a symmetry-breaking $O(2)$ Hopf bifurcation and suppose that we track this solution to finite amplitude; then how should we expect this standing wave to lose stability?

(d) In models of the Couette–Taylor apparatus where periodic boundary conditions in the axial direction are assumed, the transition from wavy vortices to modulated wavy vortices is an example of the bifurcation considered in Theorem 4.1.

Proof of Theorem 4.1. Let $x_\lambda(t)$ be the one-parameter family of standing-wave solutions to (4.1) with periods $2\pi/\omega_\lambda$. Write the Floquet equation

$$\frac{dy}{dt} = L_\lambda(t) \cdot y \tag{4.3}$$

where

$$L_\lambda(t) = (D_x F)_{x_\lambda(t),\lambda}.$$

We assume that (4.3) has a Floquet multiplier $\alpha(\lambda)$ of multiplicity two where $\alpha(0) = e^{2\pi i\theta}$ and $\theta \neq 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{2}{3}, \frac{3}{4}$. We also assume

$$\frac{d}{d\lambda}|\alpha(\lambda)|\Big|_{\lambda=0} \neq 0.$$

We know that $\dot{x}_\lambda(t)$ is always a solution to (4.3) yielding a Floquet multiplier equal to unity. Similarly, the $O(2)$ equivariance of F implies that $Jx_\lambda(t)$ is also a solution to (4.3) yielding another Floquet multiplier equal to unity where J is the infinitesimal generator of the $SO(2)$ action.

In order to eliminate this extra "trivial" Floquet multiplier, we look for solutions to (4.1) of the form

$$y(t) = R_{\eta t}x(t) \tag{4.4}$$

where R_ϕ denotes the action of $\phi \in SO(2)$ on $\mathbb{R}^N$, and η is a real parameter. As mentioned above, this trick is used in Chossat [1986] and, in a slightly different context, in Iooss [1986]. The system (4.1) now becomes

$$\dot{y} = F(y, \lambda) - \eta Jy. \tag{4.5}$$

Observe from (4.4) that periodic solutions of (4.5), $(y(t), \eta)$, correspond to quasiperiodic solutions of (4.1), $x(t)$.

Next we define our Poincaré map. Let $\phi_t(y_0, \lambda, \eta)$ denote the one-parameter group of solutions to (4.5) with initial condition y_0. Note that when $\eta = 0$, (4.5) is identical to (4.1). Recall that $x_0(t)$ is a $2\pi/\omega_0$-periodic solution to (4.1), and hence

$$x_0(0) = \phi_{2\pi/\omega_0}(x_0(0), 0, 0),$$

that is, $x_0(0)$ is a fixed point for the mapping $\phi_{2\pi/\omega_0}(\,\cdot\,, 0, 0)$.

Let $\zeta_1 = dx_0/dt(0)$ and $\zeta_2 = Jx_0(0)$. Since $x_0(t)$ is a nonconstant periodic solution to (4.1), we know that $x_0(0) \neq 0$ (since $F(0, \lambda) \equiv 0$). Thus, ζ_1 is tangent to the trajectory $x_0(t)$ and ζ_2 is tangent to the $O(2)$ group orbit through x_0. The hypothesis that $x_0(t)$ is a standing wave guarantees that ζ_1 and ζ_2 are linearly independent. Let $\langle\,,\rangle$ denote an inner product on $\mathbb{R}^N$ and let $W = \operatorname{span}\{\zeta_1, \zeta_2\}^\perp$.

We now define the first return map of trajectories to (4.5) starting in the plane $W_0 = \{x_0(0) + y_0 : y_0 \in W\}$ close to $x_0(0)$. In order for such a trajectory to return to W_0 at time τ, it must satisfy the equations

$$f_j(y_0, \lambda, \eta, \tau) \equiv \langle \zeta_j, \phi_\tau(x_0(0) + y_0, \lambda, \eta) - x_0(0)\rangle = 0. \tag{4.6}$$

Now recall that if we set $y(t) = x_0(t) + z(t)$ in (4.5) and $z(0) = y_0$ is close enough to zero, then the integral form of (4.5) is

$$z(t) = S(t)y_0 + \int_0^t S(t-s)\hat{F}(z(s), \lambda, \eta)\,ds \tag{4.7}$$

where $S(t)$ is the monodromy operator associated with (4.3) and

$$\hat{F}(z, \lambda, \eta) = F(x_0 + z, \lambda) - F(x_0, \lambda) - L_0(t)z - \eta J(z).$$

Since $\phi_\tau(x_0(0) + y_0, \lambda, \eta) = x_0(\tau) + z(\tau)$ it can be seen from (4.7) that

$$\begin{aligned} &\text{(a)} \quad f_1(y_0, \lambda, \eta, \tau) = \tau - 2\pi/\omega_0 + \cdots, \\ &\text{(b)} \quad f_2(y_0, \lambda, \eta, \tau) = \eta + \cdots, \end{aligned} \tag{4.8}$$

where $\cdots$ indicates terms of the form

$$o(|\tau - 2\pi/\omega_0| + |\eta| + O(|\lambda| + |y_0|)).$$

Using the implicit function theorem, we can solve equations (4.6) for $\tau = \tau(y_0, \lambda)$ and $\eta = \eta(y_0, \lambda)$ when $\tau(0, 0) = 2\pi/\omega_0$ and $\eta(0, 0) = 0$. Observe that generically η itself is nonzero. The Poincaré map is now defined by

$$G_\lambda(y_0) = \phi_{\tau(y_0, \lambda)}(x_0(0) + y_0, \lambda, \eta(y_0, \lambda)) - x_0(0). \tag{4.9}$$

Note that $G_0(0) = \phi_{2\pi/\omega_0}(x_0(0), 0, 0) - x_0(0) = 0$.

A consequence of this construction is that, if G_λ undergoes a Hopf bifurcation at $y_0 = 0$, then we find an invariant 2-torus in (4.5) that corresponds using (4.4) to an invariant 3-torus in (4.1). It follows from (4.4) that one of the independent frequencies is η, which is small, but typically nonzero.

A second important consequence of the construction (4.8) is that G_λ is D_n-equivariant. We claim that

$$\begin{aligned}
&\text{(a)} \quad \gamma\zeta_1 = \zeta_1 \qquad \forall \gamma \in D_n;\\
&\text{(b)} \quad \gamma\zeta_2 = \zeta_2 \qquad \forall \gamma \in \mathbb{Z}_n, \quad \text{and} \quad \phi_t(\gamma y_0, \lambda, \eta) = \gamma\phi_t(y_0, \lambda, \eta);\\
&\text{(c)} \quad S\zeta_2 = -\zeta_2 \qquad \forall S \in D_n \sim \mathbb{Z}_n, \quad \text{and} \quad \phi_t(Sy_0, \lambda, -\eta) = S\phi_t(y_0, \lambda, \eta).
\end{aligned} \tag{4.10}$$

Using the identities (4.10) in (4.6) and uniqueness of solutions to the implicit function theorem allows us to conclude that

$$\begin{aligned}
&\text{(a)} \quad \tau(\gamma y_0, \lambda) = \tau(y_0, \lambda) \qquad \forall \gamma \in D_n,\\
&\text{(b)} \quad \eta(\gamma y_0, \lambda) = \eta(y_0, \lambda) \qquad \forall \gamma \in \mathbb{Z}_n, \quad \text{and}\\
&\text{(c)} \quad \eta(Sy_0, \lambda) = -\eta(y_0, \lambda) \qquad \forall S \in D_n \sim \mathbb{Z}_n.
\end{aligned} \tag{4.11}$$

Using the definition of G_λ in (4.8), we now find it easy to check using (4.9) and (4.10) that G_λ commutes with D_n.

To verify (4.10)(a), recall that since $x_0(t)$ is a standing wave, we know that $\gamma x_0(t) = x_0(t)$ for all $\gamma \in D_n$. Now differentiate with respect to t. Next, observe that $\mathbb{Z}_n$ commutes with J while $SJ = -JS$ for all S in $D_n \sim \mathbb{Z}_n$. Now we prove (4.10)(b),(c) by invoking uniqueness of solutions to the initial value problems for systems of ODEs.

If the Floquet multipliers are simple, then this construction gives a unique invariant circle by the standard Hopf theorem for mappings. However, when the Floquet multipliers are double, we can invoke the discussion concerning Hopf bifurcation for D_n-equivariant mappings given at the end of § 3. We conclude that when G_λ undergoes a Hopf bifurcation, three families of invariant tori are produced from this bifurcation. Of course, the hypotheses of Theorem 4.1 imply that G_λ does undergo a Hopf bifurcation at $\lambda = 0$. This completes our proof. □

Remarks. (a) The stability of these 3-tori can, in principle, be computed from the results in Golubitsky and Stewart [1986b]. The simplest statement of these results is: suppose the standing waves are stable when $\lambda < 0$. Then generically, for any of the 3-tori to be stable, all those families must appear supercritically (for $\lambda > 0$). If all three families are supercritical, then precisely one family is asymptotically stable.

(b) The reader may check that these results explain the existence of the invariant 3-tori found in the interaction of two symmetry-breaking $O(2)$ Hopf bifurcations from a branch of standing-wave solutions. See Chossat, Golubitsky, and Keyfitz [1986].

5. Bifurcation from periodic solutions. In this section, we generalize the discussion of bifurcation from a periodic solution of $O(2)$ symmetric systems of ODEs to bifurcation in systems invariant under a general compact Lie group Γ. Our formulation here is mainly geometric and may be contrasted with the analytic nature of the remarks in § 4.

Let $x(t)$ be a T-periodic solution to

$$\dot{x} = f(x) \tag{5.1}$$

where $f: \mathbb{R}^n \to \mathbb{R}^n$ is Γ-equivariant, $\Gamma \subset O(n)$. Let $\phi_t: \mathbb{R}^n \to \mathbb{R}^n$ be the flow associated with (5.1) and note that ϕ_t is also Γ-equivariant. We now define a Poincaré map associated with $x(t)$.

Define a local action of $\Gamma \times \mathbb{R}$ on $\mathbb{R}^n$ by

$$(\gamma, t) \cdot x = \gamma \phi_t(x). \tag{5.2}$$

Since the actions of γ and t commute (ϕ_t is Γ-equivariant), we see that (5.2) actually defines an action of $\Gamma \times \mathbb{R}$. Let $x_0 = x(0)$. Since orbits of (smooth) Lie group actions are immersed submanifolds (cf. Bredon [1972]), we see that

$$M = (\Gamma \times \mathbb{R}) \cdot x_0 \tag{5.3}$$

is an immersed submanifold of $\mathbb{R}^n$. However, since $x(t)$ is periodic, it follows that M is compact, and hence a submanifold of $\mathbb{R}^n$. Moreover, M is foliated by the periodic solutions $\{\gamma x(t) : \gamma \in \Gamma\}$.

Remark. Let S^1 be identified with the interval $[0, T)$. Then we can define

$$\Sigma_{x_0} = \{(\gamma, \theta) \in \Gamma \times S^1 : (\gamma, \theta) x_0 = x_0\}. \tag{5.4}$$

Σ_{x_0} is the isotropy subgroup of $x(t)$ and M is diffeomorphic to $(\Gamma \times S^1)/\Sigma_{x_0}$.

Since M is compact and Γ-invariant, there exists an open Γ-invariant tubular neighborhood of M in $\mathbb{R}^n$. More precisely, there exists a vector bundle $N \overset{\pi}{\to} M$ and a smooth Γ-equivariant diffeomorphism $\sigma : N \to \mathbb{R}^n$ defined on an open neighborhood of M and N such that $\operatorname{Im} \sigma$ is an open neighborhood of M in $\mathbb{R}^n$ and $\sigma | M$ is the identity (see Bredon [1972, p. 306]). Via σ we can pull back the vector field f to N and discuss the bifurcation of the periodic orbit in N. The advantage of these coordinates is that Γ acts linearly on the fibers of N and these fibers are orthogonal to M.

Next we define the manifold

$$P_\gamma = \pi^{-1}(\{\gamma \cdot x(t)\}); \tag{5.5}$$

that is, P_γ is just that part of the vector bundle N that lies over the periodic trajectory $\gamma \cdot x(t)$ in M. It is possible to write (cf. Vanderbauwhede, Krupa, and Golubitsky [1988] or Krupa [1988])

$$f(y) = f_P(y) + f_T(y) \tag{5.6}$$

where $f_P(y)$ is tangent to P_γ for all $y \in P_\gamma$ and $f_T(y)$ is tangent to the group orbit Γy. Moreover, both f_P and f_T are Γ-equivariant.

Next, observe that Γ-equivariance implies that f_T is a linear vector field on Γy. Hence, the flow of f_T is given by $\exp(t\eta)$ for some $\eta \in L(\Gamma)$, the Lie algebra of Γ. (In fact, if we define

$$\Gamma_{x_0} = \{\gamma \in \Gamma : \gamma x_0 = x_0\} \tag{5.7}$$

and we choose a vector space complement U to $\mathcal{L}(\Gamma_{x_0})$ in $\mathcal{L}(\Gamma)$, then we can assume η is uniquely defined in U.) Note that $\mathcal{L}(O(2)) = \mathbb{R}$ and that the η defined in § 4 may be thought of as residing in the Lie algebra of $O(2)$. Also, in § 4, we solved implicitly for $\eta = \eta(y_0, \lambda)$. This discussion allows us to write explicitly the first-order terms of η. We have not set up such an explicit algorithm here. Nevertheless, we know that generically η is nonzero, which is all we need. Similarly, since f_P is Γ-equivariant, the dynamics of f_P are determined by the dynamics of $f_P | P_1$. We just transport the flow from P_1 to P_γ using multiplication by γ, which acts orthogonally.

It now follows that the flow of f may be understood as composing the flow of f_P on P_1 with linear flow on orbits Γy. In particular, if W is an invariant set under the flow of f_P, then

$$\hat{W} = \bigcup_{y \in W} \Gamma y \subset N$$

is invariant under f. Moreover, W is asymptotically stable under f_P if and only if the invariant set $\hat{W}$ is asymptotically stable under f. Since $f_T|M \equiv 0$, we may think of the flow of f on $\hat{W}$ as being the composition of the flow of f_P on W with a slow drift along the group orbits of Γ in $\hat{W}$. For the example $O(2)$, connected components of the group orbits are just circles (being diffeomorphic to $SO(2)$) and the flow for f_T along the group orbit is periodic. If, in addition, we assume that W is a 2-torus, then $\hat{W}$ will be a 3-torus with the flow along the group orbit having a small frequency.

Thus, bifurcation of the periodic orbit $x(t)$ for f is determined by bifurcation of the periodic orbit $x(t)$ for $F = f_P|P_1$. Note that, since f_P is Γ-equivariant, it follows that F is Γ_{x_0}-equivariant where Γ_{x_0} is the isotropy subgroup defined in (5.7). We assume henceforth that f, and hence F, depend on a real parameter λ.

Recall now the isotropy subgroup Σ_{x_0} of $x(t)$ in $\Gamma \times S^1$, which was defined in (5.4). We call $x(t)$ a *rotating wave* if there is a loop

$$(\gamma(\theta), \theta) \in \Sigma_{x_0}, \qquad \gamma(0) = 1 \tag{5.8}$$

and a *standing wave* otherwise. The bifurcation analysis for rotating waves proceeds along the lines of the Renardy-Rand approach. The assumption (5.8) implies that

$$x(t) = \gamma(t)^{-1} \cdot x(0).$$

Now transform the equation $\dot{y} = f_P(y)$ by looking for solutions of the form

$$y(t) = \gamma(t) z(t)$$

and obtain the system

$$\dot{z}(t) = f_P(z(t), \lambda) - \dot{\gamma}(0) z(t). \tag{5.9}$$

In this system, $x(t)$ corresponds to the steady-state solution $z(t) = x_0$ and bifurcation to tori for $x(t)$ is determined by a Hopf bifurcation in (5.9).

For standing waves, we use another approach, which is also valid for rotating waves. Let S be a cross section to $x(t)$ in the fiber of N over x_0. Since $x(t)$ is periodic, the flow of $F(\cdot, 0)$ returns to x_0 after time T. Thus, we can define the Poincaré map

$$\psi : S \times \mathbb{R} \to S$$

with $\psi(x_0, 0) = x_0$, and since F commutes with Γ_{x_0}, so does ψ. We can now study Hopf bifurcation of ψ with symmetry Γ_{x_0} using the techniques of § 3. Of course in the discussion of § 4, $\Gamma_{x_0} = D_n$ and that specific case represents an example of the general approach described here.

Acknowledgment. We are grateful to Andre Vanderbauwhede for making a number of helpful comments.

REFERENCES

V. I. ARNOLD [1977], *Loss of stability of self oscillations close to resonance and versal deformations of equivariant vector fields*, Functional Anal. Appl., 11, pp. 1-10.

———, [1983], *Geometrical Methods in the Theory of Ordinary Differential Equations*, Grundlehren 250, Springer-Verlag, New York.

E. BREDON [1972], *Introduction to Compact Transformation Groups*, Academic Press, New York.

P. CHOSSAT [1986], *Bifurcation secondaire de solutions quasi-périodiques dans un problème de bifurcation de Hopf invariant par symétrie O(2)*, C. R. Acad. Sci. Paris, 302, pp. 539-541.

P. CHOSSAT, M. GOLUBITSKY, AND B. L. KEYFITZ [1986], *Hopf-Hopf interactions with $O(2)$ symmetry*, Dynamics & Stability of Systems, 1, pp. 255-292.

G. CICOGNA [1981], *Symmetry breakdown from bifurcations*, Lett. Nuovo Cimento, 31, pp. 600-602.

T. ERNEUX AND B. J. MATKOWSKY [1984], *Quasi-periodic waves along a pulsating propagating front in a reaction-diffusion system*, SIAM J. Appl. Math., 44, pp. 536-544.

M. FIELD [1980], *Equivariant dynamical system*, Trans. Amer. Math. Soc., 259, pp. 185-205.

———, [1986], *Equivariant dynamics*, in Multiparameter Bifurcation Theory, Contemporary Mathematics 56, M. Golubitsky and J. Guckenheimer, eds., American Mathematical Society, Providence, RI, pp. 69-96.

M. GOLUBITSKY [1983], *The Bénard problem, symmetry, and the lattice of isotropy subgroups*, in Bifurcation Theory, Mechanics, and Physics, C. P. Bruter, A. Aragnol, and A. Lichnerowicz, eds., D. Reidel, Boston, MA, pp. 225-256.

M. GOLUBITSKY, J. W. SWIFT, AND E. KNOBLOCH [1984], *Symmetries and pattern selection in Rayleigh-Bérnard convection*, Physica D, 10, pp. 249-276.

M. GOLUBITSKY AND J. GUCKENHEIMER, EDS. [1986], *Multiparameter Bifurcation Theory*, Contemporary Mathematics 56, American Mathematical Society, Providence, RI.

M. GOLUBITSKY AND M. ROBERTS [1987], *Degenerate Hopf bifurcation with $O(2)$-symmetry*, J. Differential Equations, 69, pp. 216-264.

M. GOLUBITSKY AND I. STEWART [1985], *Hopf bifurcation in the presence of symmetry*, Arch. Rational Mech. Anal., 87, pp. 107-165.

———, [1986a], *Symmetry and stability in Taylor-Couette flow*, SIAM J. Math. Anal., 17, pp. 249-288.

———, [1986b], *Hopf bifurcation with dihedral group symmetry: Coupled nonlinear oscillators*, in Multiparameter Bifurcation Theory, Contemporary Mathematics 56, M. Golubitsky and J. Guckenheimer, eds., American Mathematical Society, Providence, RI, pp. 131-173.

M. GOLUBITSKY, I. STEWART, AND D. SCHAEFFER [1988], *Singularities and Groups in Bifurcation Theory: Vol.* II, Applied Mathematical Science 69, Springer-Verlag, New York.

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

———, [1984], *Bifurcation and transition to turbulence in hydrodynamics*, in CIME Session on Bifurcation Theory and Applications, L. Salvadori, ed., Lecture Notes in Mathematics 1057, Springer-Verlag, Berlin, pp. 152-201.

———, [1986], *Secondary bifurcations of the Taylor vortices into wavy inflow or outflow boundaries*, J. Fluid Mech., 173, pp. 273-288.

E. IHRIG AND M. GOLUBITSKY [1984], *Pattern selection with $O(3)$ symmetry*, Physica D, 13, pp. 1-33.

E. KNOBLOCH [1986], *On the degenerate Hopf bifurcation with $O(2)$ symmetry*, in Multiparameter Bifurcation Theory, Contemporary Mathematics 56, M. Golubitsky and J. Guckenheimer, eds., American Mathematical Society, Providence, RI, pp. 193-201.

M. KRUPA [1988], *Bifurcation of critical group orbits*, Ph.D. thesis, University of Houston, Houston, TX.

W. LANGFORD AND G. IOOSS [1980], *Interactions of the Hopf and pitchfork bifurcations*, in Bifurcation Problems and their Numerical Solutions, H. D. Mittelmann and H. Weber, eds., Birkhaüser Lecture Notes ISNM 54, Birkhäuser, Basel, pp. 103-134.

D. RAND [1982], *Dynamics and symmetry: predictions for modulated waves in rotating waves*, Arch. Rational Mech. Anal., 79, pp. 1-38.

M. RENARDY [1982], *Bifurcation from rotating waves*, Arch. Rational Mech. Anal., 75, pp. 49-84.

M. ROBERTS, J. W. SWIFT, AND D. WAGNER [1986], *The Hopf bifurcation on a hexagonal lattice*, in Multiparameter Bifurcation Theory, Contemporary Mathematics 56, M. Golubitsky and J. Guckenheimer, eds., American Mathematical Society, Providence, RI, pp. 283-318.

T. ROGERS AND D. C. WHITLEY [1983], *Chaos in the cubic mapping*, Math. Modelling, 4, pp. 9-25.

D. RUELLE [1973], *Bifurcations in the presence of a symmetry group*, Arch. Rational Mech. Anal., 51, pp. 136-152.

A. VANDERBAUWHEDE [1980], *Local Bifurcation and Symmetry*, Habilitation thesis, Rijksuniversiteit Gent. (Cf. Research Notes in Math 75, Pitman, Boston, [1982].)

A. VANDERBAUWHEDE, M. KRUPA, AND M. GOLUBITSKY [1988], *Secondary bifurcation in symmetric systems*, in Proc. of Equadiff, C. A. Dafermos, ed., 1988.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

HETEROCLINIC ORBITS AND CHAOTIC DYNAMICS IN PLANAR FLUID FLOWS*

ANDREA LOUISE BERTOZZI†

Abstract. An extension of the planar Smale–Birkhoff homoclinic theorem to the case of a heteroclinic saddle connection containing a finite number of fixed points is presented. This extension is used to find chaotic dynamics present in certain time-periodic perturbations of planar fluid models. Specifically, the Kelvin–Stuart cat's eye flow is studied, a model for a vortex pattern found in shear layers. A flow on the two-torus with Hamiltonian $H_0 = (2\pi)^{-1} \sin(2\pi x_1) \cos(2\pi x_2)$ is studied, as well as the evolution equations for an elliptical vortex in a three-dimensional strain flow.

Key words. homoclinic orbits, Melnikov's method, Kelvin–Stuart cat's eyes, elliptical vortices

AMS(MOS) subject classifications. 34C35, 54H20, 58F08, 58F13, 70K99, 76C05

1. Introduction. Organized vortex structures in two-dimensional fluid flows can often be viewed as planar dynamical systems with multiple heteroclinic saddle connections. We wish to study how such saddle connections break up under small perturbations. In the homoclinic case, the Smale–Birkhoff Theorem and Melnikov's method are two useful tools for studying the onset of chaos and mixing in planar flows possessing a simple homoclinic orbit. We extend the planar homoclinic theorem to the case of a heteroclinic orbit connecting a finite number of saddle points, enabling us to analyze fluid models to which the original homoclinic theory does not apply.

We present three planar fluid models that exhibit heteroclinic saddle connections. The Kelvin–Stuart cat's eye flow is a well-known model for a pattern found in shear layers. This flow is a planar dynamical system possessing an infinite number of heteroclinic saddle connections involving two fixed points each. We also study a planar lattice flow in which we find groups of four saddle points linked by heteroclinic orbits. The lattice flow is an interesting model for certain convection patterns as well as for nonlinear Taylor vortex flow. In the unperturbed case, these flows are steady solutions to the inviscid Euler equations and thus have a direct Hamiltonian formulation. We apply the simplified Hamiltonian form of Melnikov's method to find chaotic motion and mixing occurring in time-periodic perturbations of these two planar flows.

The third application of Melnikov's method presented here is of a somewhat different nature from the first two. We examine the evolution equations for an elliptical vortex in an imposed strain. These equations have a Hamiltonian form based on a dimensionless time parameter. The most physically interesting perturbations are based on real time and so we are forced to study a non-Hamiltonian dynamical system with a homoclinic orbit. We apply the non-Hamiltonian version of Melnikov's method to find chaotic dynamics occurring in the case of periodic stretching of the straining flow in a third dimension.

2. Extension of the homoclinic theorem and Melnikov's method. The ideas for the homoclinic theorem were first laid out by Birkhoff [5] and were developed by Smale [26]. We consider a planar diffeomorphism φ possessing a hyperbolic saddle point p whose stable and unstable manifolds intersect transversely at a point q. A result of this theorem is that φ possesses a subsystem equivalent to a shift on two symbols. We extend this theorem to the case of N fixed points joined by transverse saddle connections

* Received by the editors October 1, 1987; accepted for publication January 21, 1988.
† Department of Mathematics, Princeton University, Princeton, New Jersey 08544.

(see Fig. 2.3 for the case $N=3$). The homoclinic theorem is proved by constructing the horseshoe map and showing that it possesses the shift as a subsystem (Moser [19]). We must then show that φ possesses the horseshoe map as a subsystem. Keeping in mind Moser's proof of the homoclinic theorem, we construct the generalized horseshoe map, and present a sketch of the heteroclinic theorem. For the complete details the reader is referred to [4].

2.1. The horseshoe map and the shift on two symbols. We first define the horseshoe map used in the homoclinic case. The horseshoe map is a topological mapping of the unit square Q into the plane such that $\varphi(Q)\cap Q$ has two components U_1 and U_2. The pre-images of U_1 and U_2 are denoted by $V_i=\varphi^{-1}(U_i)$, $i=1,2$. V_1 and V_2 are vertical strips connecting the upper and lower edges of Q (see Fig. 2.1). The iterates φ^k of φ are not defined in all of Q, so we construct the invariant set

$$I=\bigcap_{k=-\infty}^{\infty}\varphi^{-k}(Q),$$

in which all iterates φ^k are defined. Associated with each point p of I is a bi-infinite sequence $(\cdots s_{-1}, s_0; s_1, s_2\cdots)$, $s_i\in\{1,2\}$ of ones and twos, where $\varphi^{-k}(p)\in V_{s_k}$ or

$$p\in\bigcap_{k=-\infty}^{\infty}\varphi^{k}(V_{s_k}).$$

On the set S of all such sequences, we define a map σ by $(\sigma s)_i=s_{i+1}$. Under the map σ, all the elements of s are shifted over by one. This provides a mapping $\tau: I\to S$ with $\tau\varphi|_I=\sigma\tau$ as long as τ is invertible. We introduce a topology on S as follows: Given $s^*=(\cdots, s^*_{-2}, s^*_{-1}, s^*_0; s^*_1, s^*_2, \cdots)\in S$ then $U_j=\{s\in S | s_k=s^*_k, (|k|<j)\}$ form a neighborhood basis for s^*. We see that the horseshoe map possesses periodic orbits of arbitrary period, as well as an orbit that comes arbitrarily close to all points of I. This last orbit is obtained by constructing a sequence that contains all possible finite strings of ones and twos.

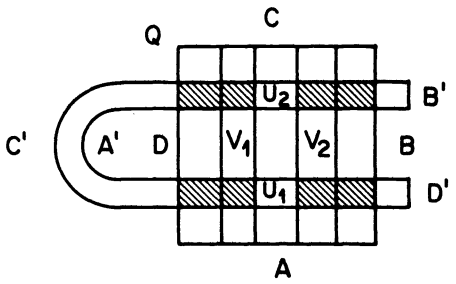


FIG. 2.1. *The horseshoe map.*

2.2. A generalization of the horseshoe map. Consider a set of N disjoint squares Q_i in the plane and a map $\varphi: \cup Q_i\to\mathbb{R}^2$ such that $\varphi(Q_i)\cap Q_i$ is a horizontal strip in Q_i and $\varphi(Q_i)\cap Q_{i+1(\text{mod } N)}$ is a horizontal strip in $Q_{i+1(\text{mod } N)}$. Here it is not important how each square Q_i is oriented with respect to the other squares, only that $\varphi(\cup_i Q_i)\cap Q_j$ are horizontal strips in Q_j (see Fig. 2.2). Our invariant set thus will be

$$I=\bigcap_{k=-\infty}^{\infty}\varphi^{-k}\left(\bigcup_i Q_i\right).$$

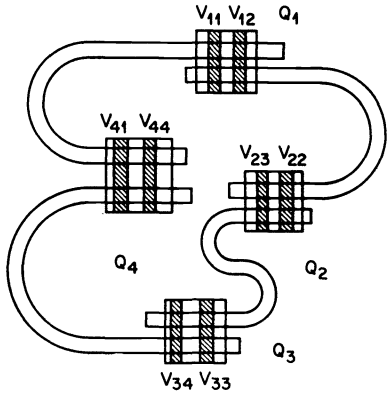


FIG. 2.2

We will associate with each point $p \in I$ a bi-infinite sequence $(\cdots, s_{-1}, s_0; s_1, s_2 \cdots) \in S'$ of N consecutive symbols where

$$S' = \{s | s_i \in \{1, \cdots, N\}, \quad s_{i+1} = s_i \quad \text{or} \quad s_{i+1} = s_i + 1 \pmod N\}$$

such that $\varphi^{-k}(p) \in Q_{s_k}$. Under the appropriate conditions there is a one-to-one correspondence between points of I and sequences $s \in S'$. For the precise details of the above construction as well as a proof of the fact that I and S' are topologically isomorphic, the reader is referred to [4].

2.3. A heteroclinic theorem.

THEOREM 2.3.1. *If a diffeomorphism $\varphi : \mathbb{R}^2 \to \mathbb{R}^2$ possesses N fixed points $p_1, p_2, \cdots, p_N$ that are nondegenerate hyperbolic saddle points, and there exist points q_i at which the unstable manifold $W^u(p_i)$ intersects the stable manifold $W^s(p_{i+1 (\mathrm{mod}\, N)})$ transversely for all i, then φ possesses an invariant set I on which some iteration φ^k is homeomorphic to the shift on S', the set of bi-infinite sequences of N consecutive symbols (as described in the preceding section).*

We provide an outline of the proof. For details, the reader is referred to [4]. We want to show that φ possesses a subsystem satisfying the requirements for the generalized horseshoe map of § 2.2. The stable and unstable manifolds are depicted in Fig. 2.3 (for the case $N = 3$).

CLAIM. We can choose an integer k and neighborhoods U_i of p_i such that the following conditions are satisfied (see Fig. 2.4):

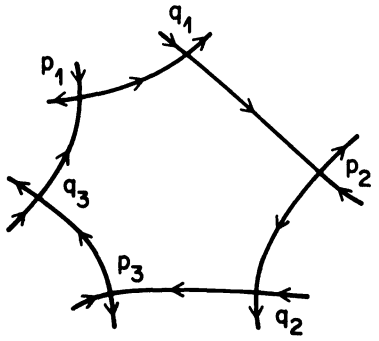


FIG. 2.3

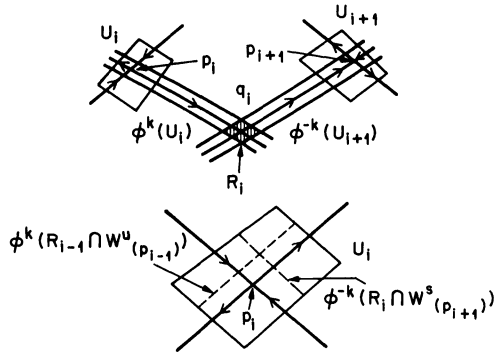


FIG. 2.4

(1) There exists a local coordinate system in U_i so that φ is linear, and U_i is the unit square.

(2) $q_i \in \varphi^k(U_i)$ and $q_i \in \varphi^{-k}(U_{i+1(\text{mod } N)})$ for all i.

(3) For $R_i = \varphi^k(U_i) \cap \varphi^{-k}(U_{i+1(\text{mod } N)})$, we have $\varphi^{-k}(R_i \cap W^s(p_{i+1(\text{mod } N)}))$ intersects $\varphi^k(R_{i-1(\text{mod } N)} \cap W^u(p_{i-1(\text{mod } N)}))$ transversely in exactly one point.

We choose U_i so that (1) is satisfied for all i. Note that if we shrink each U_i, (1) will still hold. Given any U_i satisfying (1), by the definition of stable and unstable manifolds, there exists a k such that (2) is satisfied. Note that k depends on the sizes of the U_i, which we will continue to shrink until all the above conditions are satisfied. By the λ-lemma of Palis [23], $\varphi^{-k}(R_i \cap W^s(p_{i+1(\text{mod } N)}))$ approaches $W^s(p_i)$ and $\varphi^k(R_{i-1(\text{mod } N)} \cap W^u(p_{i-1(\text{mod } N)}))$ approaches $W^u(p_i)$ as $k \to \infty$. Thus for k sufficiently large and the U_i sufficiently small, (3) is satisfied. Transversal intersection results because $W^s(p_i)$ and $W^u(p_i)$ intersect transversely at p_i. Once (3) is achieved, we can find U_i sufficiently small so that $\varphi^{-k}(R_i)$ is a vertical strip and $\varphi^k(R_{i-1(\text{mod } N)})$ is a horizontal strip in U_i. Thus, φ^{2k} possess a subsystem equivalent to the generalized horseshoe map, which in turn possesses a subsystem topologically equivalent to the shift on N consecutive symbols.

This last subsystem is termed "chaotic" because of the interesting properties it exhibits under iterations of φ^{2k}. We have orbits of arbitrary period greater than N as well as dense orbits. The bi-infinite sequence corresponding to a dense orbit is formed by concatenating all possible finite sequences of consecutive symbols. We further note the unpredictability of this subsystem. Any two orbits with sequences that agree for some finite length may have completely different sequences further on. Physically we will find these orbits near each other under a finite number of iterations of φ^{2k}, yet the orbits diverge as we proceed past the point where their sequences agree. Thus, knowing where a point will be for a fixed finite time in no way predicts where it will be at later times.

2.4. Melnikov's method. Melnikov [18] devised a method for finding the transverse intersection of stable and unstable manifolds given a time-periodic perturbation of a system with a saddle connection. We present the theorem without proof.

Consider the following planar dynamical system:

$$\text{(A)} \qquad \dot{x} = f(x) + \varepsilon g(x, t), \quad x \in \mathbb{R}^2, \quad g(x, t) = g(x, t+T), \quad 0 \leqq \varepsilon \ll 1,$$

where for $\varepsilon = 0$ we have a saddle connection Γ_0 between two nondegenerate hyperbolic saddle points p_1 and p_2 (see Fig. 2.5). The unstable manifold $W_0^u(p_1)$ of p_1 and the

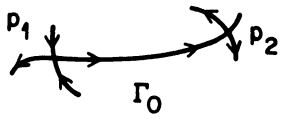


FIG. 2.5

stable manifold $W_0^s(p_2)$ of p_2 coincide. Here we include the homoclinic case where $p_1 = p_2$. Associated with (A) is the suspended system

(B) $$\dot{x} = f(x) + \varepsilon g(x, \theta), \quad (x, \theta) \in \mathbb{R}^2 \times S^1 \quad (S^1 = \mathbb{R}/T).$$

For ε sufficiently small, (B) possesses a Poincaré map: $P_\varepsilon^{t_0}: \Sigma_{t_0} \to \Sigma_{t_0}$ where $\Sigma_{t_0} = \{(x, \theta) \in \mathbb{R}^2 \times S^1 | \theta = t_0\}$ is a global cross-section of the flow. Let $F_t^\varepsilon(x_0, t_0)$ be the flow map of (B) on $\mathbb{R}^2 \times S^1$. $P_\varepsilon^{t_0}$ is obtained by a projection onto the first factor: $P_\varepsilon^{t_0}(x) = \pi(F_T^\varepsilon(x, t_0))$ where $\pi((x, \theta)) = x$. Here $P_\varepsilon^{t_0}$ is a map from $\mathbb{R}^2$ to $\mathbb{R}^2$.

Our assumptions imply that for $\varepsilon = 0$, $P_\varepsilon^{t_0}(x)$ has fixed points at p_1 and p_2 and so the suspended system has circular orbits $\gamma^1 = p_1 \times S^1$, $\gamma^2 = p_2 \times S^1$ with stable and unstable manifolds $W_0^u(\gamma^1)$ and $W_0^s(\gamma^2)$ coinciding to form a "cylinder" $\Gamma_0 \times S^1$. Such saddle connections are quite unstable and thus are expected to break under small perturbations.

We define the Melnikov function

$$M(t_0) = \int_{-\infty}^{\infty} d(q^0(t - t_0)) \wedge g(q^0(t - t_0), t) \exp\left(\int_0^{t - t_0} \operatorname{tr} Df(q^0(s))\, ds\right) dt,$$

where $q^0(t)$ is the solution to the unperturbed equation (A) starting at t_0 on the saddle connection Γ_0. We define the wedge product by $a \wedge b = a_1 b_2 - b_1 a_2$.

In the case where the unperturbed system is Hamiltonian, we have $\operatorname{tr} Df(q^0) = 0$ and the Melnikov function becomes

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

The examples of §§ 3 and 4 are both Hamiltonian systems. Two useful forms for computation are

$$M(t_0) = \int_{-\infty}^{\infty} f(q^0(t)) \wedge g(q^0(t), t + t_0) \exp\left(\int_0^t \operatorname{tr} Df(q^0(s))\, ds\right) dt$$

in the non-Hamiltonian case and

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

in the Hamiltonian case. We note that $M(t_0)$ is itself a periodic function in t_0. Using the second form, we have that

$$\begin{aligned} M(t_0 + T) &= \int_{-\infty}^{\infty} f(q^0(t)) \wedge g(q^0(t), t + t_0 + T) \exp\left(\int_0^t \operatorname{tr} Df(q^0(s))\, ds\right) dt \\ &= \int_{-\infty}^{\infty} f(q^0(t)) \wedge g(q^0(t), t + t_0) \exp\left(\int_0^t \operatorname{tr} Df(q^0(s))\, ds\right) dt \\ &= M(t_0), \end{aligned}$$

since $g(x, t + T) = g(x, t)$.

MELNIKOV'S THEOREM. *Given the above conditions, and ε sufficiently small, if $M(t_0)$ has simple zeros, then $W^s_\varepsilon(p^2_\varepsilon)$ and $W^u_\varepsilon(p^1_\varepsilon)$ intersect transversely. If $M(t_0)$ has no zeros in $t_0 \in [0, T]$ then $W^s_\varepsilon(p^2_\varepsilon) \cap W^u_\varepsilon(p^1_\varepsilon) = \varnothing$.*

For a concise proof of the homoclinic case, the reader is directed to Guckenheimer and Holmes [8]. The heteroclinic proof is an obvious generalization. For details, the reader is referred to [4].

3. Kelvin–Stuart cat's eye flow. Consider the following flow in the plane:

$$\dot{x} = \frac{a \sinh y}{a \cosh y + \sqrt{a^2-1}\cos x},$$

$$\dot{y} = \frac{\sqrt{a^2-1}\sin x}{a \cosh y + \sqrt{a^2-1}\cos x}.$$

This is a Hamiltonian system with $H_0 = \log(a \cosh y + \sqrt{a^2-1}\cos x)$. It is a model for a pattern found in shear layer flow (see [27], [12]). The parameter a controls the shape of the cat's eye with a larger a corresponding to wider "eyes." Here we consider only $a > 1$. Streamlines are constants of H_0 (see Fig. 3.1).

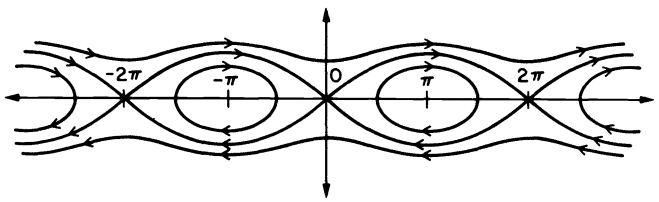


FIG. 3.1

We have fixed points at $(2\pi N, 0)$ that satisfy the conditions for Melnikov's method. Consider the upper trajectory $(x_0(t), y_0(t))$ from $(0, 0)$ to $(2\pi, 0)$. Along this trajectory we have x_0 satisfying the equation

$$\dot{x}_0 = a\sqrt{\left(\frac{a}{\sqrt{a^2-1}} + 1 - \cos x_0\right)^2 \left(\frac{a^2-1}{a^2}\right) - 1} \Big/ \left(a + \sqrt{a^2-1}\right).$$

This implicitly defines x_0 by

$$t = \int_\pi^{x_0} (a + \sqrt{a^2-1})\, dx \Big/ a\sqrt{\left(\frac{a^2-1}{a^2}\right)\left(\cos x - \frac{a}{\sqrt{a^2-1}} - 1\right)^2 - 1}.$$

By changing variables to $s = 1 - \cos x$, this integral becomes

$$\int_2^{1-\cos x_0} \left(\left(\frac{a}{\sqrt{a^2-1}} + 1\right) \Big/ s\sqrt{\left(s + \frac{2a}{\sqrt{a^2-1}}\right)(2-s)}\right) ds.$$

This can be solved exactly to yield

$$\cos x_0 = 1 - \left(\frac{8a}{a + \sqrt{a^2-1}}\right)\left(\frac{1}{e^{\gamma t} + \beta + e^{-\gamma t}}\right),$$

$$\gamma = \left(\frac{\sqrt{a^2-1}}{a + \sqrt{a^2-1}}\right)\sqrt{\frac{4a}{\sqrt{a^2-1}}}, \qquad \beta = 2\frac{a - \sqrt{a^2-1}}{a + \sqrt{a^2-1}}$$

along the upper saddle connection from $(0, 0)$ to $(2\pi, 0)$.

3.1. Periodic stretching of the cat's eye flow. Instead of examining a general perturbation $g(\vec{x}, t)$, consider a perturbation of the parameter a. If we take a to be a time-varying parameter of the form $a_0+\varepsilon b(t)$, where $b(t)$ is periodic with period T, we get a phase diagram where the "cat's eyes" are periodically stretched and compressed by an ε amount. This corresponds to a time-dependent solution to the Euler equation with external force.

To first order in ε, our perturbed equation is

$$\dot{x} = \frac{a_0 \sinh y}{a_0 \cosh y + \sqrt{a_0^2-1}\cos x} - \frac{\varepsilon b(t) \sinh y \cos x}{\sqrt{a_0^2-1}(a_0 \cosh y + \sqrt{a_0^2-1}\cos x)^2},$$

$$\dot{y} = \frac{\sqrt{a_0^2-1}\sin x}{a_0 \cosh y + \sqrt{a_0^2-1}\cos x} + \frac{\varepsilon b(t) \sin x \cosh y}{\sqrt{a_0^2-1}(a_0 \cosh y + \sqrt{a_0^2-1}\cos x)^2}.$$

Thus the driving force for our perturbation is

$$\vec{F}_\varepsilon = \frac{\varepsilon b'(t)}{\sqrt{a_0^2-1}} e^{-2H_0(x,y)} \begin{pmatrix} -\sinh y \cos x \\ \sin x \cosh y \end{pmatrix}.$$

The perturbed Hamiltonian for this system is

$$H = H_0 + \frac{\varepsilon b(t)}{\sqrt{a_0^2-1}} \left(\frac{\sqrt{a_0^2-1}\cosh y + a_0 \cos x}{a_0 \cosh y + \sqrt{a_0^2-1}\cos x} \right)$$
$$= H_0 + H_1.$$

Along all streamlines of the unperturbed flow,

$$H_1 \propto b(t)\left(\sqrt{a_0^2-1}\cosh y + a_0 \cos x\right).$$

Since the saddle connections are streamlines of the unperturbed flow, how they break up under a perturbation depends only on the perturbation at the points of the saddle connection. Thus, the Melnikov function for the above perturbation is identical to the one corresponding to the simpler perturbation

$$H_1 = \varepsilon b(t)\left(\sqrt{a^2-1}\cosh y + a \cos x\right).$$

If we let $b(t)$ have the form $\cos(kt)$, then this perturbation corresponds to the superposition of four waves:

$$\sqrt{a^2-1}\cosh y(e^{i(z-kt)} + e^{i(z+kt)}) + a(e^{i(x-kt)} + e^{i(x+kt)}).$$

Here z is the third coordinate and we take the cross-sectional flow in the plane $z=0$. The wavelength of the perturbation is exactly equal to the length of one of the cat's eyes. The wave speed is allowed to vary.

3.2. The Melnikov function for periodic stretching. Consider the upper trajectory $(x_0(t), y_0(t))$ from $(0, 0)$ to $(2\pi, 0)$ for the unperturbed system.

The Melnikov function for this trajectory is

$$M(t_0) = \int_{-\infty}^{\infty} C_1[(a_0 \sin x_0(t) \cosh y_0(t) \sinh y_0(t)$$
$$+\sqrt{a_0^2-1}\sinh y_0(t) \cos x_0(t) \sin x_0(t) b(t+t_0))]\, dt,$$

which can be reduced to

$$M(t_0) = \int_{-\infty}^{\infty} C_2(\sin x_0(t) \sinh y_0(t) b(t+t_0))\, dt$$

where

$$C_2 = \frac{1}{\sqrt{a_0^2-1}(a_0+\sqrt{a_0^2-1})^2}.$$

Here we have exploited the fact that

$$a_o \cosh y_0(t) + \sqrt{a_0^2-1} \cos x_0(t) = a_0 + \sqrt{a_0^2-1}.$$

We expand $b(t)$ into its Fourier series:

$$b(t) = \sum_{-\infty}^{\infty} (a_k \sin kt + b_k \cos kt).$$

The above Melnikov integral then becomes

$$\sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} C_2 \sin x_0(t) \sinh y_0(t)(a_k \sin k(t+t_0) + b_k \cos k(t+t_0))\, dt$$

$$= \sum_{k=-\infty}^{\infty} \left((a_k \cos kt_0 - b_k \sin kt_0) \int_{-\infty}^{\infty} C_2 \sin x_0(t) \sinh y_0(t) \sin (kt)\, dt \right)$$

$$= \sum_{k=-\infty}^{\infty} ((a_k \cos (kt_0) - b_k \sin (kt_0)) M_0(k)),$$

where we define

$$M_0(k) = \int_{-\infty}^{\infty} C_3 \frac{d}{dt}[\cos x_0(t)] \sin (kt)\, dt,$$

$$C_3 = -(a_0+\sqrt{a_0^2-1})C_2.$$

We have used the fact that $\sin x_0(t) \sinh y_0(t)$ is an odd function in t. Thus,

$$M_0(k) = \int_{-\infty}^{\infty} C_4 \frac{e^{\gamma t} - e^{-\gamma t}}{(e^{\gamma t} + \beta_0 + e^{-\gamma t})^2} \sin (kt)\, dt,$$

$$C_4 = C_3\left(\gamma\left(\frac{8a_0}{a_0+\sqrt{a_0^2-1}}\right)\right).$$

Evaluation by residues (see Appendix A) yields, for $k \neq 0$,

$$\left(\frac{\gamma}{2\pi C_4}\right) M_0(k) = \left[\frac{m e^{-|m|\alpha}}{2 \sin \alpha} - \left(\frac{e^{-|m|2\pi}}{1-e^{-|m|2\pi}}\right)\left(\frac{m \sinh |m|\alpha}{\sin \alpha}\right)\right],$$

$$m = \frac{k}{\gamma}, \quad \alpha = \cos^{-1}\left(\frac{\beta_0}{2}\right), \quad 0 < \alpha < \pi/2.$$

Whether or not $M(t_0)$ has simple zeros depends on the values of a_k and b_k. For instance, if $b(t)$ is of the form $\cos kt$, then we see that $M(t_0)$ has simple zeros for almost all k. A similar analysis shows that the lower trajectory has a Melnikov function that is just the negative of the one for the upper trajectory. Since both trajectories break up under the same perturbation to yield the transverse intersection of stable and unstable manifolds, we have satisfied the requirements for the heteroclinic theorem (Theorem 2.3.1) with $N = 2$. Our perturbed system has a chaotic subsystem topologically equivalent to a shift on two symbols.

3.3. Mixing in the perturbed cat's eye flow. By exploiting the symmetry of this model, we see that this perturbing function breaks up all trajectories transversely. In fact, we can view both the perturbed and unperturbed cases as flows on the cylinder. Here we take $x \in \mathbb{R}/2\pi$, $y \in \mathbb{R}$. All of the saddle points are identified and we obtain two homoclinic orbits to a single saddle point. We can now use the standard homoclinic theorem to find a shift on two symbols.

Based on the proof of the theorem from the second section, we expect mixing to occur at least within the region around the fixed point. We know that there exists a neighborhood U of the fixed point $(0, 2\pi N)$ on which the Poincaré map for this system acts like a version of the horseshoe map (see Fig. 3.2).

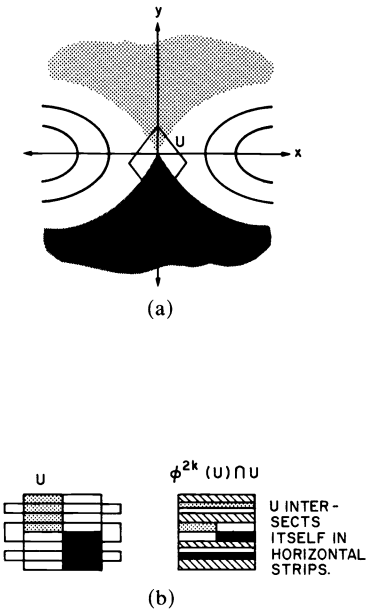


FIG. 3.2. (a) *The cat's eye flow on the cylinder.* (b) *Perturbed cat's eye flow. Here the top and bottom layers are mixed into the cat's eyes region and eventually into each other.*

Viewed as a flow on the plane, we see that the perturbed system has a geometric structure similar to that of Holmes's perturbed sine-Gordon equation [11, § 3]. We show that the perturbed cat's eye flow has a subsystem isomorphic to the shift on the symbols "+" and "−," where the "+" corresponds to traveling "downstream" along an upper trajectory and the "−" corresponds to traveling "upstream" along a lower trajectory (see Fig. 3.3). This provides a mechanism for fluid inside one cat's eye to travel both upstream and downstream. This mechanism does not exist for the unperturbed case, since flow within an "eye" will remain there for all time. In the perturbed system, all saddle connections are broken up to give us transversal intersection of stable and unstable manifolds. The heteroclinic theorem tells us that at each fixed point $p_n = (2\pi n, 0)$, there is a neighborhood U_n, a unit square in local coordinates, such that for some fixed time T^*, the flow φ_{T^*} maps U_i to intersect U_{i-1} and U_{i+1} in horizontal strips. A simplified model of the dynamics present is pictured in Fig. 3.4. Here each U_i is intersected by the horizontal strips $H_{i-1,i} = \varphi(U_{i-1}) \cap U_i$ and $H_{i+1,i} = \varphi(U_{i+1}) \cap U_i$.

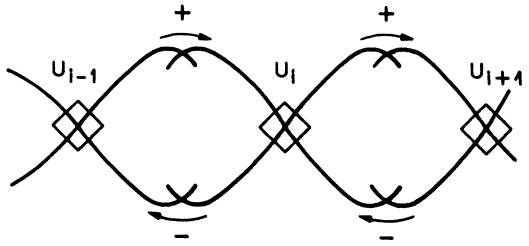


FIG. 3.3

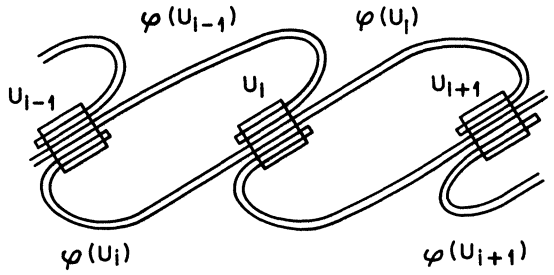


FIG. 3.4

By the symmetry of the flow and its perturbation, we can choose each U_i so that $U_i + 2\pi = U_{i+1}$ and $\varphi(U_i) + 2\pi = \varphi(U_{i+1})$. Our invariant set is

$$I = \bigcap_{k=-\infty}^{\infty} \varphi^{-k}\left(\bigcup_i (H_{i+1,i} \cup H_{i-1,i})\right).$$

I can be decomposed into disjoint sets $I_i = U_i \cap I$. For any given i, we have a one-to-one correspondence between I_i and $S^{\pm}$, the set of all bi-infinite sequences of "+" and "−":

$$\tau : I_i \to S^{\pm},$$
$$[\tau(x)]_l = \quad + \quad \text{if } \varphi^l(x) \in U_k \Rightarrow \varphi^{l+1}(x) \in U_{k+1},$$
$$= \quad - \quad \text{if } \varphi^l(x) \in U_k \Rightarrow \varphi^{l+1}(x) \in U_{k-1}.$$

Thus there is a set $S^{\pm}$ of sequences corresponding to each I_i. We see that there is a mechanism for pieces of fluid to move rather chaotically both upstream and downstream as well as for fluid within each "eye" to mix with fluid in other "eyes." This mixing and chaotic motion was not present in the unperturbed cat's eye flow. The fact that the perturbation cos kt leads to such chaos for almost all k indicates that such mixing may be rather common in the actual shear layers.

4. Planar lattice flow. We consider the following flow:

$$\dot{x}_1 = -\sin(2\pi x_1)\sin(2\pi x_2), \qquad \dot{x}_2 = -\cos(2\pi x_1)\cos(2\pi x_2)$$

a Hamiltonian system with $H_0 = (2\pi)^{-1}\sin(2\pi x_1)\cos(2\pi x_2)$ (see Fig. 4.1). This is a model for axisymmetric Taylor vortex flow as well as for many convective flows. If we take x_1 to be a moving coordinate, these equations model the Rossby waves of geophysical fluid dynamics (see [24, p. 84]). This flow is obviously doubly periodic, yielding a flow on the torus $T = \mathbb{R}^2/\Gamma$ where Γ is the lattice $\{(n_1, n_2);\ n_1, n_2 \in \mathbb{Z}\}$.

Viewed as a flow on the torus T, we obtain a system with heteroclinic orbits connecting four saddle points. Melnikov's theory can then be applied to perturbations of this flow.

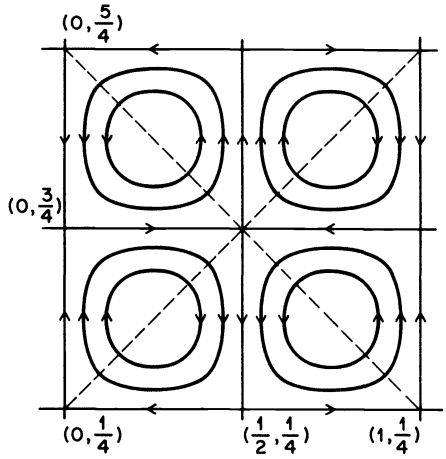


FIG. 4.1. Γ' *is represented by the dashed line.*

We can also map this flow onto a "smaller" torus $T' = \mathbb{R}^2/\Gamma'$ where $\Gamma' = \{(\frac{1}{2}(n_1 - n_2), \frac{1}{2}(n_1 + n_2))\}$ (see Fig. 4.1). Here we have exploited the periodicity in the variables $(x_1 - x_2)$, $(x_1 + x_2)$ as well as in x_1 and x_2. The flow on T' has only two heteroclinic saddle points. By examining perturbed flows on T', we can look for a subsystem that is a shift on two symbols. This horseshoe-like structure will result if all heteroclinic orbits are broken up so that stable and unstable manifolds intersect transversely.

4.1. Time- and space-dependent perturbations. We consider two types of perturbations, ones that are functions of time only and ones that have an added space dependence. In the purely time-dependent case, we have $\varepsilon \vec{f}(t)$ as a perturbation to the velocity field, with $\vec{f}(t) = \vec{f}(t+T)$. This corresponds to an external driving force $\vec{F}_\varepsilon = \varepsilon \vec{f}'(t)$ that is uniform in space at any given moment. This is physically reasonable as an approximation to an external force that is time-periodic and has an average space variation much larger than the periodic lattice structure of the flow. For the vertical saddle connections, the Melnikov function for this perturbation is

$$M_v(t_0) = \pm \int_{-\infty}^{\infty} \cos(2\pi x_2(t)) f_1(t + t_0)\, dt$$

since $\sin(2\pi x_1) = 0$ for these trajectories. Likewise for the horizontal orbits, $\cos(2\pi x_2) = 0$ and so

$$M_h(t_0) = \pm \int_{-\infty}^{\infty} \sin(2\pi x_1(t)) f_2(t + t_0)\, dt.$$

We see that the vertical and horizontal components of $\vec{f}$ are decoupled. We will show by symmetry properties that f_1 and f_2 must satisfy the same conditions in order for M_v and M_h to have simple zeros. For this space-independent perturbation, the Γ'-lattice symmetry is preserved and chaotic motion can be reduced to a subsystem isomorphic to the shift on two symbols. The following example presents a spatially dependent perturbation that breaks up the Γ' symmetry.

In general, a perturbing velocity of the form

$$\varepsilon \begin{pmatrix} v_1(x_2, t) \\ v_2(x_1, t) \end{pmatrix}$$

constitutes a solution to the two-dimensional Euler equation with external force

$$\vec{F}_\varepsilon = \varepsilon \begin{pmatrix} \partial v_1(x_2, t)/\partial t \\ \partial v_2(x_1, t)/\partial t \end{pmatrix}.$$

A particularly interesting perturbation of this form is

$$\varepsilon \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \varepsilon \begin{pmatrix} \cos(2\pi x_2)\cos kt \\ \sin(2\pi x_1)\cos kt \end{pmatrix}.$$

This has a stream function

$$\frac{\varepsilon}{2\pi}\cos kt[\sin(2\pi x_2) - \cos(2\pi x_1)],$$

which can be viewed as a superposition of linear waves traveling along coordinate axes:

$$-(e^{i(2\pi x_1 + kt)} + e^{i(2\pi x_1 - kt)}) - i(e^{i(2\pi x_2 + kt)} - e^{i(2\pi x_2 - kt)}).$$

This perturbation is geometrically interesting because it breaks up the Γ' symmetry and we are forced to consider heteroclinic orbits joining four points instead of two points. We shall show that for almost all k, the saddle connections break up to yield a subsystem topologically equivalent to the shift on four consecutive symbols.

4.2. Explicit calculation of the Melnikov functions. Along an unperturbed horizontal saddle connection, we have

$$\dot{x}_1 = \pm\sin(2\pi x_1), \qquad \dot{x}_2 = 0$$

and along a vertical connection

$$\dot{x}_2 = \pm\cos(2\pi x_2), \qquad \dot{x}_1 = 0.$$

In the case of the connection from $(\frac{1}{2}, \frac{1}{4})$ to $(0, \frac{1}{4})$, we have $\dot{x}_1 = -\sin(2\pi x_1)$. This has a solution $x_1 = (1/\pi)\tan^{-1}(e^{-2\pi t})$, which by symmetry properties of the flow yields

$$\sin(2\pi x_1) = \pm\frac{2e^{-2\pi t}}{1+e^{-4\pi t}}$$

along all horizontal connections and

$$\cos(2\pi x_2) = \pm\frac{2e^{-2\pi t}}{1+e^{-4\pi t}}$$

along all vertical ones.

For a spatially independent perturbation, the Melnikov function of § 4.1, for either saddle connection, is of the form

$$M_i(t_0) = \int_{-\infty}^{\infty} \frac{2e^{-2\pi t}}{1+e^{-4\pi t}} f_i(t+t_0)\, dt.$$

If we expand f_i into its Fourier series

$$f_i(t) = \sum_{k=-\infty}^{\infty} \left(A_k \cos\left(\frac{2\pi kt}{T}\right) + B_k \sin\left(\frac{2\pi kt}{T}\right)\right)$$

we find that

$$M_i(t_0) = \sum_{k=-\infty}^{\infty} \left[\left(A_k \cos\left(\frac{2\pi kt_0}{T}\right) + B_k \sin\left(\frac{2\pi kt_0}{T}\right)\right) \int_{-\infty}^{\infty} \frac{2e^{-2\pi t}}{1+e^{-4\pi t}} \cos\left(\frac{2\pi kt}{T}\right) dt \right].$$

Evaluation by residues reveals

$$M_i(t_0) = \sum_{k=-\infty}^{\infty} \left(A_k \cos\left(\frac{2\pi k t_0}{T}\right) + B_k \sin\left(\frac{2\pi k t_0}{T}\right)\right)\left(\frac{1}{e^{-\pi k/2T} + e^{\pi k/2T}}\right).$$

Whether or not $M_i(t_0)$ has simple zeros depends on the respective values of A_k and B_k. For example, if $f_i = A_0 + A_1 \cos(2\pi k t/T)$, we require

$$|A_0| < |A_1| \left[\frac{2}{e^{-\pi k/2T} + e^{\pi k/2T}}\right]$$

for $M_i(t_0)$ to have simple zeros. Now we see that the class of perturbing functions $\vec{f} = (A \cos(t), B \sin(t))$ yields $M_v(t_0)$ and $M_h(t_0)$ with simple zeros for all saddle connections. Applying the results of § 2, we obtain a shift on four symbols as a subsystem of the perturbed flow on T, and a shift on two symbols as a subsystem of the perturbed flow on T'.

For the spatially dependent perturbation

$$\varepsilon \begin{pmatrix} \cos(2\pi x_2) \cos kt \\ \sin(2\pi x_1) \cos kt \end{pmatrix}$$

we find that, up to a change of sign, the Melnikov function for either a vertical or horizontal saddle connection is

$$M(t_0) = 4 \cos(kt_0) \int_{-\infty}^{\infty} \frac{e^{-4\pi t}}{(1+e^{-4\pi t})^2} \cos kt \, dt$$

which we evaluate via residues (using the procedure outlined in Appendix A for the calculation of the integral in § 3) to be

$$\begin{aligned} M(t_0) &= \cos kt_0 \left(\frac{k}{4\pi \sinh(k/4)}\right) \\ &= \cos kt_0 M_0(k). \end{aligned}$$

$M_0(k)$ is nonzero for almost all k so that the Melnikov function will have simple zeros and we have a subsystem topologically equivalent to the shift on four consecutive symbols.

4.3. Mixing in the perturbed lattice flow. Under both perturbations, we expect some sort of mixing to occur that was not present in the unperturbed case. In the perturbed systems, all connections are broken up to yield transverse intersection of stable and unstable manifolds. As in the cat's eye model, at each fixed point $p_{n_1 n_2} = (\frac{1}{2}n_1, \frac{1}{2}n_2 + \frac{1}{4})$, $n_1, n_2 \in \mathbb{Z}$, we have neighborhoods $U_{n_1 n_2}$ that intersect each other in horizontal strips under some fixed time mapping of the flow (see Fig. 4.2). In the case of the first perturbation studied, we can exploit the T' symmetry to obtain a subsystem topologically equivalent to the shift on two symbols. The perturbed and unperturbed systems are both flows on the torus T'. Under this symmetry, we can identify all clockwise rotating cells with each other and likewise all counterclockwise rotating cells with each other (Fig. 4.3). In the unperturbed case, these patches of fluid do not mix. The perturbation satisfies the conditions of the heteroclinic theorem with two fixed points, yielding a subsystem of the flow topologically equivalent to the shift on two symbols. In the perturbed case we see mixing patterns similar to those present in the cat's eye flow.

In the case of the second perturbation, we do not have the T' symmetry. The cells break up into two different clockwise and counterclockwise rotations (see Fig. 4.3).

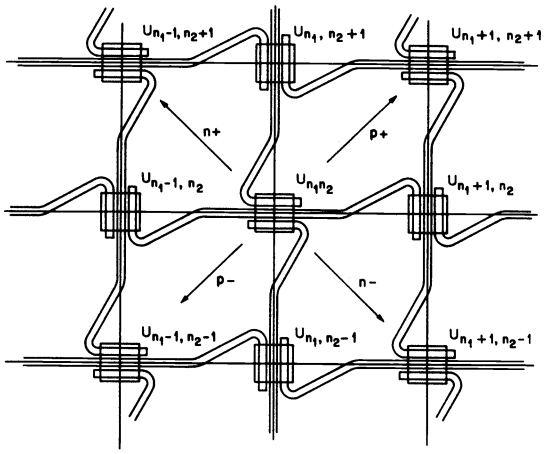
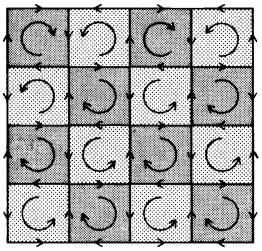
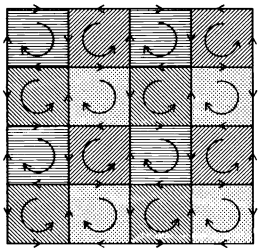


FIG. 4.2



(a)



(b)

FIG. 4.3. (a) *Flow on T'. All clockwise rotating cells are identified, as are all counterclockwise rotating cells.* (b) *Flow on T. There are two types of clockwise rotations as well as two types of counterclockwise rotations.*

On the torus T we have four fixed points in the heteroclinic orbit and our system breaks up to yield a subsystem topologically equivalent to the shift on four consecutive symbols. In the previous case we have symbols 1 and 2 identified with 3 and 4, respectively. This is analogous to identifying the two clockwise rotations with each other and likewise the two counterclockwise rotations with each other. Again we expect similar mixing patterns to occur.

In the cat's eye flow, we found a mechanism for traveling up- and downstream randomly within the cat's eyes. This corresponded to a shift in the symbols "+" and "−." In the lattice flow, we find a mechanism for traveling all over the plane, along the Γ' lattice. We find that the perturbed lattice flow has a subsystem isomorphic to the shift on the four symbols "n_+," "n_-," "p_+," "p_-." Here, $n_\pm$ corresponds to a

translation by $\pm(-\frac{1}{2}, \frac{1}{2})$ along the lattice. Likewise $p_\pm$ corresponds to a translation by $\pm(\frac{1}{2}, \frac{1}{2})$ (see Fig. 4.2).

In the neighborhood $U_{n_1 n_2}$ of each fixed point $p_{n_1 n_2}$, we find that for some fixed time T^*, the flow φ_{T^*} maps $U_{n_1 n_2}$ to intersect U_{n_1-1,n_2} and U_{n_1+1,n_2} for n_1+n_2 odd, or U_{n_1,n_2-1} and U_{n_1,n_2+1} for n_1+n_2 even, in horizontal strips (see Fig. 4.2).

These strips are mapped to smaller strips in U_{n_1-1,n_2-1}, U_{n_1+1,n_2-1}, U_{n_1-1,n_2+1}, U_{n_1+1,n_2+1}, by a second iteration of φ_{T^*}. Thus φ_{2T^*} maps $U_{n_1 n_2}$ to intersect U_{n_1-1,n_2-1}, U_{n_1+1,n_2-1}, U_{n_1-1,n_2+1}, U_{n_1+1,n_2+1}, in horizontal strips.

For convenience, we now refer to φ_{2T^*} as φ. Thus our invariant set is

$$I = \bigcap_{k=-\infty}^{\infty} \varphi^{-k}\left(\bigcup_{n_1,n_2} (\varphi(U_{n_1 n_2}) \cap (U_{n_1-1,n_2-1} \cup U_{n_1+1,n_2-1} \cup U_{n_1-1,n_2+1} \cup U_{n_1+1,n_2+1}))\right),$$

where I can be decomposed into the disjoint sets $I_{n_1 n_2} = U_{n_1 n_2} \cap I$. For any given pair (n_1, n_2), there is a one-to-one correspondence between $I_{n_1 n_2}$ and the set of bi-infinite sequences on the symbols p_+, p_-, n_+, n_-:

$$\tau : I_{n_1 n_2} \to S$$

$$[\tau(x)]_l = p_+ \quad \text{if } \varphi^l(x) \in U_{n_1 n_2} \Rightarrow \varphi^{l+1}(x) \in U_{n_1+1,n_2+1},$$

$$[\tau(x)]_l = p_- \quad \text{if } \varphi^l(x) \in U_{n_1 n_2} \Rightarrow \varphi^{l+1}(x) \in U_{n_1-1,n_2-1},$$

$$[\tau(x)]_l = n_+ \quad \text{if } \varphi^l(x) \in U_{n_1 n_2} \Rightarrow \varphi^{l+1}(x) \in U_{n_1-1,n_2+1},$$

$$[\tau(x)]_l = n_- \quad \text{if } \varphi^l(x) \in U_{n_1 n_2} \Rightarrow \varphi^{l+1}(x) \in U_{n_1+1,n_2-1}.$$

Thus, fluid particles within one cell can travel randomly around the plane in the perturbed case. In the unperturbed case, this sort of mixing is not allowed since fluid within one cell will remain there for all time.

5. Motion of an elliptical vortex in a strain field. An important part of fluid mechanics is the study of vortices, their structure, and how they interact with one another. In §§ 3 and 4 we examined two well-known two-dimensional planar fluid models. Since organized vortex structures are observed frequently, we would like to find a simple model for a vortex affected by a field of neighboring vortices. As the examples of §§ 3 and 4 indicate, the presence of multiple vortices in stationary planar fluid flow often results in fixed points of the flow, between vortex structures, that can be modeled as hyperbolic saddle points in a planar dynamical system. In a neighborhood of such saddle points, the velocity field is roughly linear and can be locally approximated by a simple strain. Thus it is physically reasonable to model certain vortex interaction locally as a single vortex in a straining flow. Moore and Saffman [20], as well as Neu [21], describe vortex interaction that can be modeled in such a way.

We study the motion of an elliptical vortex in a three-dimensional imposed strain. We see that the evolution of such a vortex can be characterized as a planar dynamical system that has interesting Hamiltonian and non-Hamiltonian formulations involving the aspect ratio $\eta = a/b$ and the angle θ of rotation of the ellipse. Here a and b correspond to the major and minor axes of the ellipse. We apply Melnikov's method to the evolution equations of the vortex to show chaotic dynamics occurring in the presence of three-dimensional periodic stretching of the imposed strain. The actual analysis differs somewhat from what was done in the previous sections in that we study chaos occurring in the evolution equation of the shape and orientation of the ellipse as opposed to chaos occurring in the flow pattern of an actual fluid model.

5.1. Hamiltonian formulation of exact Euler solution. The Hamiltonian formulation presented below is due to Neu [22] and represents a three-dimensional generalization of the exact solutions of an elliptical vortex in a two-dimensional straining flow (described by Kida [14]). First consider a planar vortex region in the shape of an ellipse with constant vorticity in the interior. The points on the boundary of the region are solutions to the equation $x^2/a^2+y^2/b^2=\text{constant}$. Following a potential theory calculation described in Lamb [15], we see that the velocity field inside the ellipse is linear:

$$\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix}=\tilde{U}(a,b,\theta)\begin{pmatrix}x\\ y\end{pmatrix},$$

$$\tilde{U}(a,b,\theta)=-\frac{\omega}{a+b}R(\theta)\begin{pmatrix}0 & a\\ -b & 0\end{pmatrix}R(\theta).$$

Here a and b correspond, respectively, to the major and minor axes of this elliptical cross-section and θ is the angle of the major axis with respect to the x-axis. $R(\theta)$ is the rotation matrix

$$\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}.$$

In three dimensions we have a cylindrical vortex region whose cross-section in the xy-plane is the above velocity field. We add an irrotational straining field the velocity of which is given by $v=(\gamma' x,-\gamma y,\gamma'' z)$ where $\gamma'-\gamma+\gamma''=0$ is required for incompressibility. The combination of vortex and strain yields a fluid velocity that, in the xy-plane, has the form $U(a,b,\theta)(x,y)^T$ where

$$U(a,b,\theta)=-\frac{\omega}{a+b}R(\theta)\begin{pmatrix}0 & a\\ -b & 0\end{pmatrix}R(\theta)+\begin{pmatrix}\gamma' & 0\\ 0 & -\gamma\end{pmatrix}.$$

The velocity field inside the vortex is again linear and the path of a fluid particle on the boundary must satisfy the equation of an ellipse which we write in matrix form:

$$\begin{pmatrix}x\\ y\end{pmatrix}E(a,b,\theta)\begin{pmatrix}x & y\end{pmatrix}=\text{constant},$$

$$E(a,b,\theta)=R(\theta)\begin{pmatrix}a^{-2} & 0\\ 0 & b^{-2}\end{pmatrix}R(\theta).$$

Differentiating the ellipse equation, we obtain

$$\dot{X}^TEX+X^T\dot{E}X+X^TE\dot{X}=0$$

where X is the vector (x,y). Since $\dot{X}=U(a,b,\theta)X$, we have the matrix evolution equation

$$\dot{E}+U^TE+EU=0,$$

which we can write out explicitly in terms of a, b, and θ to give us the evolution equations for the elliptical vortex:

$$\dot{a}+(\gamma\sin^2\theta-\gamma'\cos^2\theta)a=0,$$

$$\dot{b}+(\gamma\cos^2\theta-\gamma'\sin^2\theta)b=0,$$

$$\dot{\theta}=\frac{\omega ab}{(a+b)^2}-\frac{1}{2}(\gamma+\gamma')\frac{a^2+b^2}{a^2-b^2}\sin 2\theta.$$

These evolution equations have the following Hamiltonian formulation: let $\eta=a/b$ be the aspect ratio and τ be a dimensionless time defined by $d\tau/dt=\omega\eta^2/(\eta^2-1)$.

Then the evolution equations become

$$\frac{d\eta}{d\tau} = -\frac{\partial H}{\partial \theta} = \frac{\gamma+\gamma'}{\omega}\left(\eta - \frac{1}{\eta}\right)\cos 2\theta,$$

$$\frac{d\theta}{d\tau} = \frac{\partial H}{\partial \eta} = \frac{\eta-1}{\eta(1+\eta)} - \frac{1}{2}\frac{\gamma+\gamma'}{\omega}\left(1+\frac{1}{\eta^2}\right)\sin 2\theta,$$

$$H = \log\frac{(1+\eta)^2}{\eta} - \frac{1}{2}\frac{\gamma+\gamma'}{\omega}\left(\eta - \frac{1}{\eta}\right)\sin 2\theta.$$

We consider γ, γ', and ω to be, in general, time-dependent parameters in this equation. The total circulation of the vortex is $\Gamma = \pi ab\omega$, which we know to be constant by the Kelvin Circulation Theorem (see [6, p. 28]). The evolution equations imply that

$$\frac{d(ab)}{dt} = (\gamma' - \gamma)ab$$

so that

$$ab = a_0 b_0 e^{(\gamma'-\gamma)t},$$

which in turn yields

$$\omega = \omega_0 e^{(\gamma-\gamma')t}.$$

Thus, when $\gamma'' = 0$, $\gamma' - \gamma + \gamma'' = 0$ implies that $\gamma = \gamma'$. Our Hamiltonian system is autonomous if and only if $\gamma'' = 0$, γ, γ' are both constant. We will consider the case where this autonomous Hamiltonian system is perturbed by a periodic stretching of the strain where we set $\gamma'' = \varepsilon g(t)$.

In the autonomous case, we have $\gamma = \gamma'$, and are interested in the dynamics indicated in the phase portrait for $0 < \gamma/\omega < 0.15$ (see Fig. 5.1). There are no heteroclinic

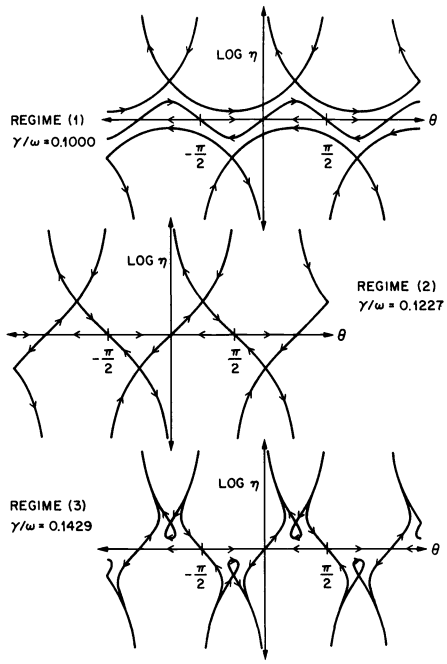


FIG. 5.1

orbits in the phase portrait for $\gamma/\omega > .15$. The three interesting regimes are depicted in Fig. 5.1:

(1) For $0 < \gamma/\omega < .1227$, there are oscillating regions (bubbles close to the $\log \eta = 0$ axis) as well as rotating regions between the bubbles and the outer saddle connections.

(2) At $\gamma/\omega = .1227$ we have a bifurcation where saddle connections between three fixed points exist for this value of γ/ω only.

(3) For γ/ω between .1227 and .15, we have homoclinic saddle connections, the interior of which represents an ellipse oscillating about the ray $\theta = \pi/4$.

The importance of the bifurcation is that in regimes (2) and (3) we no longer have the possibility of a rotating ellipse.

5.2. Real time formulation of evolution equations. In order to apply Melnikov's method to the above Hamiltonian system, we would need to consider time-periodic perturbations of the dimensionless time τ. This is not a reasonable physical model, since a periodic perturbation of the straining flow would be periodic in real time, and not in the dimensionless time τ. Note that the evolution equations written in terms of the orientation, aspect ratio and real time are

$$\frac{d\theta}{dt} = \frac{\omega\eta}{(\eta+1)^2} - \frac{1}{2}(\gamma+\gamma')\frac{\eta^2+1}{\eta^2-1}\sin 2\theta,$$

$$\frac{d\eta}{dt} = (\gamma+\gamma')\eta\cos 2\theta.$$

Since (η, θ) and $(\eta^{-1}, \theta+\pi/2)$ correspond to the same ellipse, we can parameterize the evolution equation in terms of $r = \log \eta$, $\varphi = 2\theta$, and yield a polar coordinates formulation for these equations in which there is a one-to-one correspondence between ellipses and points in the phase space (r, φ). The evolution equations become

$$\dot{r} = (\gamma+\gamma')\cos\varphi$$

$$\dot{\varphi} = \frac{2\omega e^r}{(e^r+1)^2} - (\gamma+\gamma')\frac{e^{2r}+1}{e^{2r}-1}\sin\varphi.$$

From the Hamiltonian formulation, we know that trajectories correspond to constants of

$$H = \log\left[\frac{(1+e^r)^2}{e^r}\right] - \frac{\gamma+\gamma'}{2\omega}(e^r - e^{-r})\sin\varphi.$$

This can be verified by calculating $dH/dt = 0$ for the real time t. These equations seem to blow up for $r = 0$. Fortunately, we see that this blow-up is due to the coordinates we are using and not the equations themselves. Polar coordinates are not well defined at $r = 0$ so we convert the equations to Cartesian form by $x = r\cos\varphi$, $y = r\sin\varphi$. The evolution equations become:

$$\dot{x} = (\gamma+\gamma')\frac{x^2}{x^2+y^2} - \frac{2\omega y e^{\sqrt{x^2+y^2}}}{(e^{\sqrt{x^2+y^2}}+1)^2} + (\gamma+\gamma')\frac{e^{2\sqrt{x^2+y^2}}+1}{e^{2\sqrt{x^2+y^2}}-1}\frac{y^2}{\sqrt{x^2+y^2}},$$

$$\dot{y} = (\gamma+\gamma')\frac{xy}{x^2+y^2} + \frac{2\omega x e^{\sqrt{x^2+y^2}}}{(e^{\sqrt{x^2+y^2}}+1)^2} - (\gamma+\gamma')\frac{e^{2\sqrt{x^2+y^2}}+1}{e^{2\sqrt{x^2+y^2}}-1}\frac{xy}{\sqrt{x^2+y^2}}.$$

We see that as $r \to 0$, the first and third terms in $\dot{x}$ appear to blow up. Using Taylor

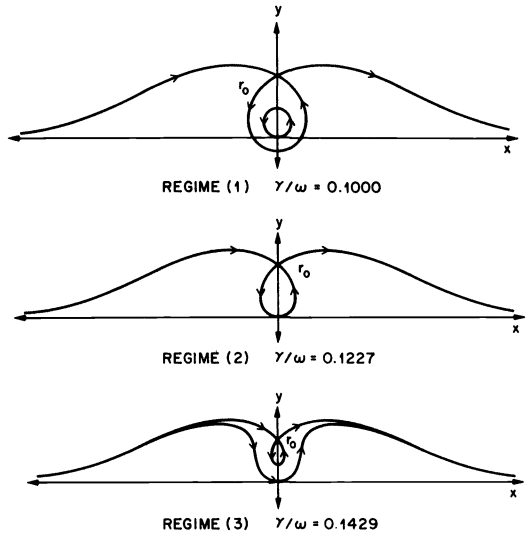


FIG. 5.2

expansion techniques, we see that the third term can be approximated by

$$(\gamma+\gamma')\frac{y^2}{x^2+y^2}(1+\mathcal{O}(r^2))$$

for r small. Thus, $\dot{x}\to\gamma+\gamma'$ as $r\to 0$. In a similar fashion, we see that $\dot{y}\to 0$ as $r\to 0$.

The phase portrait (Fig. 5.2) for the real time formulation has a much simpler form than that of the Hamiltonian one we first introduced (Fig. 5.1). We see that for $(\gamma+\gamma')/2\omega<.15$, there is a homoclinic loop with hyperbolic fixed point corresponding to the largest root of $e^r(e^r-1)=((\gamma+\gamma')/2\omega)(e^{2r}+1)(e^r+1)$. We see that the bifurcation at $\gamma/\omega=.1227$ is represented by the loop crossing the origin.

5.3. Periodic stretching of an elliptical vortex. In general, our perturbed system will have the form

$$\dot{r}=C_0\cos\varphi+\varepsilon g_1(r,\varphi,t),$$

$$\dot{\varphi}=\frac{2\omega_0 e^r}{(e^r+1)^2}-C_0\frac{e^{2r}+1}{e^{2r}-1}\sin\varphi+\varepsilon g_2(r,\varphi,t).$$

Here g_1 and g_2 are periodic in time, $C_0=(\gamma_0+\gamma_0')$.

For $0<C_0/\omega_0<0.15$, the unperturbed system has a hyperbolic fixed point p_0 at $\varphi=\pi/2$, $r=r_0$ where r_0 corresponds to the largest real root of the cubic $e^{3r}+e^{2r}(1-B)+e^r(1+B)+1=0$ where $B=(\frac{1}{2}C_0\omega_0)^{-1}$. This fixed point has a homoclinic saddle connection Γ_0 as depicted in Fig. 5.2. If we consider a perturbation involving a periodic stretching by an amount $\varepsilon\gamma''(t)$, then our perturbation has the form

$$g_1=C_1(t)\cos\varphi,$$

$$g_2=\frac{2C_2(t)e^r}{(e^r+1)^2}-C_1(t)\frac{e^{2r}+1}{e^{2r}-1}\sin\varphi.$$

Here C_1 and C_2 are periodic in time with period T. We consider the symmetric case where the oscillation of γ'' puts equal and opposite oscillations on γ' and γ while

maintaining the incompressibility condition $\gamma' - \gamma + \gamma'' = 0$. Thus, $\gamma + \gamma'$ stays constant even though $\gamma - \gamma'$ oscillates with γ''. This implies $C_1(t) = 0$ so that our perturbation has the simpler form

$$g_1 = 0, \qquad g_2 = \frac{2C_2(t)e^r}{(e^r+1)^2}.$$

If we parameterize Γ_0 by $(r(t), \varphi(t))$, the Melnikov function for the perturbed system can be calculated using the non-Hamiltonian form. There are two ways of doing the Melnikov function calculation. We can view Γ_0 as a trajectory in the (r, φ) coordinate system, which has the advantage of a simpler formulation. Since these coordinates break up at $r = 0$, we cannot treat the case where Γ_0 contains the point $r = 0$. This occurs only at the value $\gamma/\omega = .1227$. For any other value of γ/ω, we can find a C^∞ vector field $(f_1(r, \varphi), f_2(r, \varphi))$ so that

$$\dot{r} = f_1(r, \varphi), \qquad \dot{\varphi} = f_2(r, \varphi)$$

is a planar differentiable dynamical system in the coordinates (r, φ) with a saddle connection identical to Γ_0 in its real time parameterization. We have $f_1 = C_0 \cos \varphi$, $f_2 = 2\omega_0 e^r/(e^r+1)^2 - C_0 \sin \varphi (e^{2r}+1)/(e^{2r}-1)$ in a neighborhood of the curve Γ_0. This new dynamical system is suitable for Melnikov's method and in a neighborhood of Γ_0 has dynamics identical to that of the original system.

Alternatively, we can treat the evolution equation as a dynamical system in the (x, y) coordinates. This allows us to show that chaos will also occur in the degenerate case of $\gamma/\omega = .1227$. Both calculations are presented.

The Melnikov function in (r, φ) *coordinates.* For this we need to know $\exp\left(\int_0^t \operatorname{tr} Df(\Gamma_0(s))\, ds\right)$. We have

$$\operatorname{tr} Df = -C_0 \frac{e^{2r}+1}{e^{2r}-1} \cos \varphi = \frac{-\dot{r}(e^{2r}+1)}{e^{2r}-1}.$$

This gives us

$$\exp\left(\int_0^t \operatorname{tr} Df(\Gamma_0(s))\, ds\right) = \frac{e^{r(t)}(e^{r_0} - e^{-r_0})}{e^{2r(t)}-1}.$$

This yields a Melnikov function

$$M(t_0) = C_3 \int_{-\infty}^{\infty} \frac{e^{2r} \cos \varphi}{(e^r+1)^2(e^{2r}-1)} C_2(t+t_0)\, dt,$$
$$C_3 = C_0(e^{r_0} - e^{-r_0}).$$

Using the fact that the integral represents a convolution with an odd function, for $C_2 = \cos kt$, we have

$$M(t_0) = C_3 \sin kt_0 \int_{-\infty}^{\infty} \frac{e^{2r} \cos \varphi}{(e^r+1)^2(e^{2r}-1)} \sin kt\, dt$$
$$= \sin kt_0 M_0(k).$$

Since $\cos \varphi \propto \dot{r}(t)$, we can see that the above integral is the sine transform of an L^1 function.

$$\int_{-\infty}^{\infty} \left| \frac{e^{2r} \cos \varphi}{(e^r+1)^2(e^{2r}-1)} \right| dt = 2 \int_0^{\infty} \left| \frac{e^{2r} \cos \varphi}{(e^r+1)^2(e^{2r}-1)} \right| dt$$
$$\leqq \frac{C_3}{C_0} \int_{r(0)}^{r(\infty)} \frac{e^{2r}}{(e^r+1)^2(e^{2r}-1)}\, dr < \infty.$$

We know by the properties of the Fourier transform on $L^1(\mathbb{R})$ ([13, pp. 120–131]) that $M_0(k)$ is a uniformly continuous function of k that is not identically zero. Thus there exists some interval $k_1 \leqq k \leqq k_2$ such that $M_0(k)$ is nonzero. For these values of k, $M(t_0)$ has simple zeros.

The Melnikov function in (x, y) coordinates. We now consider the dynamical system

$$\dot{x} = (\gamma+\gamma')\cos^2\varphi - \frac{2\omega r \sin\varphi e^r}{(e^r+1)^2} + (\gamma+\gamma')\frac{e^{2r}+1}{e^{2r}-1} r\sin^2\varphi, \qquad r \neq 0,$$

$$\dot{y} = (\gamma+\gamma')\cos\varphi\sin\varphi + \frac{2\omega r\cos\varphi e^r}{(e^r+1)^2} - (\gamma+\gamma')\frac{e^{2r}+1}{e^{2r}-1} r\sin\varphi\cos\varphi, \qquad r \neq 0,$$

$$\dot{x} = \gamma+\gamma', \qquad \dot{y} = 0,$$

for $r = 0$. Here $r = \sqrt{x^2+y^2}$, $\varphi = \tan^{-1}(y/x)$. We have the time-periodic perturbation

$$\vec{g}(t, x, y) = \frac{\varepsilon C_2(t) r e^r}{(e^r+1)^2}\begin{pmatrix} -\sin\varphi \\ \cos\varphi \end{pmatrix}.$$

The following analysis is for the case $\gamma+\gamma' = .1227$; the nondegenerate case can be studied in a similar fashion. We have

$$f \wedge g = (\gamma+\gamma')\cos\varphi C_2(t_0)\left(\frac{re^r}{(e^r+1)^2}\right),$$

$$\operatorname{tr} Df = (\gamma+\gamma')\cos\varphi\left(\frac{1}{r} - \frac{e^{2r}+1}{e^{2r}-1}\right) \qquad r \neq 0,$$

$$= 0, \qquad r = 0,$$

$$e^{\int_0^t \operatorname{tr} Df ds} = \frac{2r(t)e^{r(t)}}{e^{2r(t)}-1}.$$

For $C_2(t) = \cos kt$, our Melnikov function is

$$M(t_0) = -\int_{-\infty}^{\infty} \sin kt_0 \dot{r}(t) \sin kt \frac{r^2 e^{2r}}{(e^{2r}-1)(e^r+1)^2}\, dt$$

$$= \sin kt_0 M_0(k).$$

Again we see that $M_0(k)$ is a sine transform of an L^1 function:

$$\int_{-\infty}^{\infty} \left|\frac{\dot{r}(t) r^2 e^{2r}}{(e^{2r}-1)(e^r+1)^2}\right| dt = \int_{r(0)}^{r(\infty)} \frac{r^2 e^{2r}}{(e^{2r}-1)(e^r+1)^2}\, dr$$

$$= \int_0^{r(\infty)} \frac{r^2 e^{2r}}{(e^{2r}-1)(e^r+1)^2}\, dr$$

$$< \infty.$$

This is because $r^2 e^{2r}/(e^{2r}-1)(e^r+1)^2$ is bounded on the interval $(0, r(\infty)]$. We see that $M_0(k)$ is again the sine transform of an odd L^1 function so that there exists an interval $k_1 \leqq k \leqq k_2$ so that $M_0(k)$ is nonzero, giving us a Melnikov function with simple zeros.

Under such a periodic stretching, we find chaotic dynamics occurring in the phase portrait of the evolution equations for the ellipse. This indicates a sort of randomness in the evolution of the vortex. The phase portrait includes a horseshoe as a subsystem that, as we know from § 2, indicates somewhat erratic behavior on an invariant set. Assuming the inability to make completely precise measurements, we can only predict what will happen to the vortex for a finite time; after this time we have no knowledge of how it will evolve.

Appendix A. We present the details of the calculation of the following integral from § 3 via residues:

$$M_0(K) = \int_{-\infty}^{\infty} \frac{e^{\gamma t} - e^{-\gamma t}}{(e^{\gamma t} + \beta + e^{-\gamma t})^2} \sin kt \, dt,$$

which by a change of variables $\tau = \gamma t$ becomes

$$M_0'(m) = \frac{1}{\gamma} \int_{-\infty}^{\infty} \frac{e^{\tau} - e^{-\tau}}{(e^{\tau} + \beta + e^{-\tau})^2} \sin m\tau \, d\tau,$$

where $m = k/\gamma$. Consider the meromorphic function

$$\frac{(e^{3z} - e^{z}) e^{imz}}{(e^{2z} - \beta e^{z} + 1)^2}.$$

The denominator has roots

$$e^{z} = \frac{\beta \pm \sqrt{\beta^2 - 4}}{2},$$

which we can write as

$$= e^{\pm i\alpha},$$

since we know that $0 < \beta < 2$. Here, $\alpha = \cos^{-1}(\beta/2)$, which gives us $0 < \alpha < \pi/2$. Thus, the function

$$\frac{(e^{3z} - e^{z}) e^{imz}}{(e^{z} - e^{i\alpha})^2 (e^{z} - e^{-i\alpha})^2}$$

has double poles at $z = \pm i\alpha + 2\pi i N$, $N \in \mathbb{Z}$. Let $r = z - (i\alpha + 2\pi i N)$. The integral is clearly odd in m. Thus we need only consider the case $m > 0$. We have that

$$\begin{aligned}
&\operatorname{Im} \frac{1}{\gamma} \int_{-\infty}^{\infty} \frac{e^{3\tau} - e^{\tau}}{(e^{2\tau} + \beta e^{\tau} + 1)^2} e^{im\tau} \, d\tau \\
&\quad = \lim_{N \to \infty} \operatorname{Im} \frac{1}{\gamma} \int_{-(2N+1)\pi}^{(2N+1)\pi} \frac{e^{3\tau} - e^{\tau}}{(e^{2\tau} + \beta e^{\tau} + 1)^2} e^{im\tau} \, d\tau \\
&\quad = \lim_{N \to \infty} \operatorname{Im} \frac{1}{\gamma} \left[\frac{1}{2\pi i} \sum_{0 < y < (2N+1)\pi} \operatorname{Res} \left(\frac{e^{3z} - e^{z}}{(e^{2z} + \beta e^{z} + 1)^2} e^{imz} \right) \right. \\
&\qquad \left. - \int_{|z| = (2N+1)\pi, y > 0} \frac{e^{3\tau} - e^{\tau}}{(e^{2\tau} + \beta e^{\tau} + 1)^2} e^{im\tau} \, d\tau \right].
\end{aligned}$$

The last integral goes to zero as $N \to \infty$ so that for $m > 0$, we wish to calculate

$$\operatorname{Im} \left[\frac{2\pi i}{\gamma} \sum_{y > 0} \operatorname{Res} \left(\frac{e^{3z} - e^{z}}{(e^{2z} + \beta e^{z} + 1)^2} e^{imz} \right) \right].$$

Thus we need to calculate the residues of the function in the upper half-plane. We can calculate the coefficients of the Laurent expansion of the function by first considering the expansion of its components in the neighborhood of $i\alpha+2\pi iN$. Writing $r=z-(i\alpha+2\pi iN)$, we have

$$e^{3z}=e^{3i\alpha}\left(1+3r+\frac{9}{2}r^2+\cdots\right),$$

$$e^{z}=e^{i\alpha}\left(1+r+\frac{1}{2}r^2+\cdots\right),$$

$$e^{imz}=e^{-m\alpha-2\pi mN}\left(1+imr-\frac{m^2}{2}r^2+\cdots\right),$$

$$(e^z-e^{i\alpha})^2=e^{2i\alpha}(r^2+r^3+\cdots),$$

$$(e^z-e^{-i\alpha})^2=-4\sin^2\alpha+4i\sin\alpha e^{i\alpha}r+\cdots.$$

We write the function in the form

$$\frac{1}{r^2}\left(\frac{a+br+\cdots}{c+dr+\cdots}\right),$$

which has the Laurent expansion

$$\frac{a}{c}r^{-2}+\left(\frac{b}{c}-\frac{da}{c^2}\right)r^{-1}+\cdots,$$

so that the residue at $r=0$ is $b/c-da/c^2$. Here

$$a=(e^{3i\alpha}-e^{i\alpha})e^{-m\alpha-2\pi mN},$$

$$b=e^{-m\alpha-2\pi mN}[im(e^{3i\alpha}-e^{i\alpha})+3e^{3i\alpha}-e^{i\alpha}],$$

$$c=e^{2i\alpha}(-4\sin^2\alpha),$$

$$d=e^{2i\alpha}(4ie^{i\alpha}\sin\alpha-4\sin^2\alpha).$$

Let R_x denote the residue at the point x. Then,

$$R_{i\alpha+2\pi iN}=\frac{me^{-m\alpha-2\pi mN}}{2\sin\alpha}.$$

Notice that $R_{i\alpha+2\pi iN}=R_{i\alpha}e^{-2\pi Nm}$. A similar calculation shows that

$$R_{-i\alpha+2\pi iN}=\frac{-me^{m\alpha-2\pi mN}}{2\sin\alpha}.$$

We add the residues in the upper half plane to obtain $M_0'(m)$ for $m>0$, and exploit the fact that $M_0'(m)$ is odd in m to obtain $M_0'(m)$ for $m<0$. Thus for $m\neq 0$, our integral becomes

$$M_0'(m)=\frac{2\pi}{\gamma}\left[\frac{me^{-|m|\alpha}}{2\sin\alpha}-\frac{m\sinh|m|\alpha}{\sin\alpha}\frac{e^{-2\pi|m|}}{1-e^{-2\pi|m|}}\right].$$

Acknowledgments. I would like to credit Prof. Andrew Majda of Princeton University for originally suggesting the idea of extending the Melnikov theory to study the onset of chaos in fluid flows with heteroclinic orbits. I would also like to thank him for the guidance he has given me on my A.B. thesis, from which this paper stems. I would like to thank AT&T Bell Laboratories at Murray Hill, New Jersey, where I spent the summers of 1987 and 1988, for supplying me with the diagrams.

REFERENCES

[1] L. V. AHLFORS, *Complex Analysis*, McGraw-Hill, New York, 1979.

[2] V. I. ARNOLD, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1984.

[3] V. I. ARNOLD AND A. AVEZ, *Ergodic Problems of Classical Mechanics*, W. A. Benjamin, New York, Amsterdam, 1968.

[4] A. L. BERTOZZI, *An extension of the Smale-Birkhoff homoclinic theorem, Melnikov's method, and chaotic dynamics in incompressible fluids*, A.B. thesis, Princeton University, Princeton, NJ, 1987.

[5] G. D. BIRKHOFF, *Nouvelles recherches sur les systèmes dynamiques*, Mem. Pont Acad. Sci. Novi Lyncaei, 1 (1935), pp. 85-216.

[6] A. J. CHORIN AND J. E. MARSDEN, *A Mathematical Introduction to Fluid Mechanics*, Springer-Verlag, New York, 1979.

[7] J. GUCKENHEIMER, *A Brief Introduction to Dynamical Systems*, Lectures in Applied Mathematics 17, Springer-Verlag, Berlin, New York, 1979, pp. 187-252.

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

[9] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[10] P. J. HOLMES, *Averaging and chaotic motions in forced oscillations*, SIAM J. Appl. Math., 38 (1980), pp. 68-80.

[11] ———, "*Space and Time Periodic Perturbations of the sine-Gordon Equation*, in Dynamical Systems and Turbulence, D. A. Rand and L.-S. Young, eds., Lecture Notes in Mathematics 89, Springer-Verlag, New York, Berlin, 1981.

[12] HOLM, MARSDEN, AND RATIU, *Nonlinear stability of the Kelvin-Stuart cat's eyes flow, in* AMS Proc. Math. Bio. Symposia and Summer Seminars, July, 1979.

[13] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, Dover, New York, 1976.

[14] S. KIDA, J. Phys. Soc. Japan, 50 (1981), pp. 3517-3520.

[15] H. LAMB, *Hydrodynamics*, Dover, New York, 1945.

[16] A. MAJDA, Lectures on Incompressible Fluid Flow-Fall 1986, Princeton, NJ, 1986.

[17] J. E. MARSDEN, *Chaotic orbits by Melnikov's method: a survey of applications*, PAM-173/COMM, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, August 1983.

[18] V. K. MELNIKOV, *On the stability of the center for time periodic perturbations*, Trans. Moscow Math. Soc., 12 (1963), pp. 1-57.

[19] J. MOSER, *Stable and Random Motions in Dynamical Systems*, Princeton University Press, Princeton, NJ, 1973.

[20] D. W. MOORE AND P. G. SAFFMAN, *The density of organized vortices in a turbulent mixing layer*, J. Fluid Mech., 69 (1975), pp. 465-473.

[21] J. C. NEU, *The Dynamics of a Columnar Vortex in an Imposed Strain*, MRSI 022-83, 1983.

[22] ———, *The dynamics of stretched vortices*, J. Fluid Mech., 143 (1984), pp. 253-276.

[23] J. PALIS, *On Morse-Smale dynamical systems*, Topology, 8 (1969), pp. 385-405.

[24] J. PEDLOSKY, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1982.

[25] A. P. PRUDNIKOV, Y. A. BRYCHKOV, AND O. I. MARICHEV, *Integrals and Series: Vol. I. Elementary Functions* (translated from Russian by N. M. Queen) Gordon and Breach, New York, 1986.

[26] S. SMALE, *Differomorphisms with many periodic points*, in Differential and Combinatorial Topology, S. S. Cairns, ed., Princeton University Press, Princeton, NJ, 1963, pp. 63-80.

[27] J. T. STUART, *Stability Problems in Fluids*, AMS Lectures in Applied Mathematics 13, American Mathematical Society, Providence, RI, 1971, pp. 139-155.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

003

BIFURCATIONS OF RELATIVE EQUILIBRIA IN THE N-BODY AND KIRCHHOFF PROBLEMS*

KENNETH R. MEYER† AND DIETER S. SCHMIDT‡

Abstract. The bifurcations of a one-parameter family of relative equilibria in the N-body problem are studied using normal form theory, Lie transforms, and an algebraic processor. The one-parameter family consists of $N-1$ bodies of mass 1 at the vertices of a regular polygon and one body of mass m at the centroid. As N increases there are more and more values of the mass parameter m where the relative equilibrium is degenerate. For $N \leqq 13$ each of these degenerates gives rise to a bifurcation and a new relative equilibrium. This is established using a computer-aided proof. A similar analysis is carried out for the N-vortex problem of Kirchhoff.

Key words. central configurations, relative equilibria, N-body, bifurcation

AMS(MOS) subject classification. 70F10

1. Introduction. The study of relative equilibria (r.e.) of the N-body problem has had a long history starting with the famous collinear configuration of the 3-body problem found by Euler (1767). Over the intervening years many different technologies have been applied to the study of r.e. In the older papers of Euler (1767), Lagrange (1772), Hoppe (1879), Lehmann-Filhes (1891), and Moulton (1910), special coordinates, symmetries, and analytic techniques were used. In their investigations, Dziobek (1900) used the theory of determinants; Smale (1970) used Morse theory; Palmore (1975) used homology theory; Simo (1977) used a computer; and Moeckel (1985) used real algebraic geometry. Thus, the study of r.e. has been a testing ground for many different methodologies of mathematics.

In Meyer and Schmidt (1987) the methods of bifurcation analysis and the use of the automated algebraic processor were brought to bear on this subject and the present paper continues the attack. Specifically we study the bifurcations of the relative equilibrium which consists of $N-1$ particles of mass 1 at the vertices of a regular polygon and one particle of mass m at the centroid. We call this the regular polygon relative equilibrium (r.p.r.e.). Our first paper considers the 4- and 5-body problems and uses the special coordinates of Dziobek (1900). These coordinates make the 4-body problem relatively easy to handle and the 5-body problem accessible, but beyond 5, Dziobek's coordinates become very cumbersome. The 4- and 5-body problems in these special coordinates are sufficiently simple that the general purpose algebraic processor MACSYMA could handle the tedious calculations. For larger N the special purpose algebraic processor POLYPAK written by the second author was needed because the computations increased rapidly with N. In the analysis of the 4- and 5-body problems the classical power series methods of bifurcation analysis handles the problems nicely, but for larger n a systematic use of Lie transforms by Deprit (1969) was mandated in order to bring the equations into a normal form. Thus this paper uses substantially different techniques than our previous paper.

* Received by the editors May 15, 1987; accepted for publication (in revised form) January 6, 1988. This research was supported by a grant from the Applied and Computational Mathematics Program of the Defense Advanced Research Projects Agency.

† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221-0025.

‡ Department of Computer Science, University of Cincinnati, Cincinnati, Ohio 45221-0008.

The problem of finding an r.e. can be reduced to finding a critical point of the potential energy function on the manifold of constant moment of inertia. Thus the problem falls within the domain of catastrophe theory and so the general theory is well understood. However, this specific problem has a high degree of symmetry, many variables, and a constraint, so the computations must be performed with care. We consider this paper as a case study in bifurcation analysis in the face of these complexities.

Indeed we analyze the problem at three different computational levels. First, for small N, we perform the normalization to high order to determine the existence, uniqueness, and exact shape of the bifurcating equilibria. For medium ranges of N we exploit the symmetry so that fewer computations need be carried out in order to establish existence, but now the uniqueness is only within the class of symmetric equilibria. Last, for large N, we carry out some calculations to establish existence of bifurcations with no uniqueness information. We can see that for a fixed amount of computing power the precision of the information obtained decreases as N increases.

For the planar problem that we consider, a relative equilibrium is also a central configuration and vice versa, that is, a homothetic solution which begins or ends in total collapse or tends to infinity. Even though as solutions of the N-body problem r.e. are quite rare and rather special, they are of central importance in the analysis of the asymptotic behavior of the universe. In general, solutions which expand beyond bounds or collapse in a collision do so asymptotically to a central configuration. A survey and entrance to this literature can be found in Saari (1980).

Interestingly this problem in celestial mechanics is formally similar to the problem in fluid dynamics of describing the evolution of finitely many interacting point vortices in the plane. Kirchhoff (1897) shows that this problem is specified by a Hamiltonian which is similar to the Hamiltonian of an N-body problem with a logarithmic potential. The constants that correspond to the masses are now the circulations, which may be positive or negative, and so a richer store of bifurcations are to be expected. We develop the theory and evolution of the bifurcations of the problem in parallel with that of the N-body problem.

In Meyer and Schmidt (1987) we studied the 4- and 5-body problem and found that there was a unique value of the mass of the central particle where the potential was degenerate. This agrees with the findings in Palmore (1973). However, for larger N there are more and more values of this mass at which the potential is degenerate, which disagrees with Palmore (1976). In fact, for large N many bifurcations occur. We developed the general theory of the bifurcations for these two problems for all N and completely analyze the bifurcations for $4 \leqq N \leqq 13$. Figures 1 and 2 illustrate the bifurcations which occur at the unique critical mass when $N = 4, 5$ and Fig. 3 illustrates the multitude of bifurcations that occurs in the 13-body problem.

Also we found that the self-potential for the N-body problem with the central mass removed was not always a nondegenerate minimum. In fact it is a saddle for $N > 6$. This disagrees with one of the findings in Palmore (1975). There were other surprises in our investigations, which will be explained below when we have developed the necessary definitions and notation.

2. Relative equilibria for the N-body and Kirchhoff problems. The N-body problem is the system of differential equations that describes the motion of N particles moving under the influence of their mutual gravitational attraction. Let $q_j \in R^2$ be the position vector, $p_j \in R^2$ the momentum vector, and $m_j > 0$ the mass of the jth particle, $1 \leqq j \leqq N$;

then the equations of motion are

$$\begin{aligned} \dot{q}_j &= \frac{\partial H}{\partial p_j} = \frac{1}{m_j} p_j, \\ \dot{p}_j &= -\frac{\partial H}{\partial q_j} = \frac{\partial U}{\partial q_j}, \end{aligned} \qquad j = 1, \cdots, N \tag{2.1}$$

where H is the Hamiltonian

$$H = \sum_{j=1}^{N} \frac{\|p_j\|^2}{2m_j} - U(q) \tag{2.2}$$

and U is the (self) potential

$$U = \sum_{1 \leqq i < j \leqq N} \frac{m_i m_j}{\|q_i - q_j\|}. \tag{2.3}$$

These equations reduce to the Newtonian formulation

$$m_j \ddot{q}_j = \frac{\partial U}{\partial q_j}, \qquad j = 1, \cdots, N. \tag{2.4}$$

To change to rotating coordinates let $q_j = \exp(\nu J t) u_j$ where $\nu > 0$ is the frequency of the rotating frame and

$$J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

so (2.4) becomes

$$m_j \{\ddot{u}_j + 2\nu J \dot{u}_j - \nu^2 u_j\} = \frac{\partial U}{\partial u_j}(u), \qquad j = 1, \cdots, N. \tag{2.5}$$

An equilibrium in these rotating coordinates is a solution of the system of algebraic equations

$$-\lambda m_j u_j = \frac{\partial U}{\partial u_j}, \qquad j = 1, \cdots, N \tag{2.6}$$

where $\lambda = \nu^2 > 0$.

The Kirchhoff problem is the system of differential equations describing the motion of N vortices moving in the plane under their mutual interaction. Let q_j be the position vector and $m_j \neq 0$ the circulation of the jth vortex for $j = 1, \cdots, N$. Then Kirchhoff (1897) gives the equations of motion as

$$m_j \dot{q}_j = J \frac{\partial U(q)}{\partial q_j}, \qquad j = 1, \cdots, N \tag{2.7}$$

where now U is the Hamiltonian

$$U = -\sum_{1 \leqq i < j \leqq N} m_i m_j \log \|q_i - q_j\|. \tag{2.8}$$

Introducing rotating coordinates as before by setting $q_j = \exp(\nu Jt)u_j$ transforms (7) to the system

$$m_j\{\dot{u}_j + \nu J u_j\} = J\frac{\partial U(u)}{\partial u_j}, \qquad j = 1, \cdots, N. \tag{2.9}$$

An equilibrium in these rotating coordinates is a solution of the system of algebraic equations

$$-\lambda m_j u_j = \frac{\partial U(u)}{\partial u_j}, \qquad j = 1, \cdots, N \tag{2.10}$$

where $\lambda = -\nu$.

It is classical and easy to verify that if $\bar{u} = (\bar{u}_1, \cdots, \bar{u}_N)$ and $\bar{\lambda}$ is a solution of (6) (or (10), respectively), then the center of mass of $\bar{u}$ is at the origin ($\sum m_j \bar{u}_j = 0$) and $\bar{\lambda} = U(\bar{u})/(1+\delta)I(\bar{u})$ (>0 for (6)) where I is the moment of inertia

$$I(u) = \tfrac{1}{2}\sum m_j \|u_j\|^2 \tag{2.11}$$

and $\delta = 0$ for the Kirchhoff problem or $\delta = 1$ for the N-body problem. For either problem we will set

$$\begin{aligned} M &= \{u \in R^{2N} : \textstyle\sum m_j u_j = 0\}, \\ \Delta &= \{u \in R^{2N} : u_i = u_j \text{ for some } i \neq j\}, \\ S &= \{u \in M : I(u) = 1\}. \end{aligned} \tag{2.12}$$

The variable λ can be considered a Lagrange multiplier and so an equivalent definition of a relative equilibrium is a critical point of U restricted to $S\backslash\Delta$. If u is an r.e. then so is $Au = (Au_1, \cdots, Au_N)$ where $A \in \mathrm{SO}(2, R)$ is a rotation matrix. We can define an equivalence relation by $u \sim Au$ when $A \in \mathrm{SO}(2, R)$, and since U, I, are constant on equivalence classes we can define the quotient spaces $\mathcal{S} = (S\backslash\Delta)/\sim$ and the function $\mathcal{U} : \mathcal{S} \to R$ by $\mathcal{U}([u]) = U(u)$, where [ ] denotes an equivalence class. $\mathcal{S}$ and $\mathcal{U}$ are smooth. Thus a similarity class of r.e. is a critical point of $\mathcal{U}$.

A relative equilibrium is called nondegenerate if its equivalence class is a nondegenerate critical point of $\mathcal{U}$ in the sense of Morse theory, i.e., the Hessian is nonsingular at the critical point. It follows from the implicit function theorem that bifurcations can occur only at degenerate critical points, so first we must find degenerate r.e.

3. Palmore coordinates. Our first step is to introduce the local coordinate system on the quotient space $\mathcal{S}$ which was given in Palmore (1976). Let $n = N-1$ and $\omega = \exp(i2\pi/n)$ be a primitive nth root of unity. Consider complex numbers as vectors in the plane, so ω^j, $0 \leqq j < n$ are the vertices of a regular polygon with n sides. By the regular polygon relative equilibrium (r.p.r.e.) we shall mean the r.e. which consists of n particles of unit mass, $m_j = 1$, situated at ω^j for $j = 0, \cdots, n-1$, and one particle of arbitrary mass, $m_n = m$, situated at the origin.

Let $q = (q_0, q_1, \cdots, q_n)^T$ be the position vector of the $N = n+1$ particles in the plane, $\Omega = (\omega^0, \omega^1, \omega^2, \cdots, \omega^{n-1}, 0)^T$ be the position vector of the r.p.r.e., and change coordinates by

$$q = \Omega + \mathcal{V}z \tag{3.1}$$

where $z=(z_0, z_1, \cdots, z_n)$ is the position vector in the new Palmore coordinates and $\mathcal{V}$ is the matrix

$$\mathcal{V}=\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & \omega^1 & \omega^2 & \cdots & \omega^{n-1} & 1 \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-1)} & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)^2} & 1 \\ 0 & 0 & 0 & \cdots & 0 & -n/m \end{pmatrix}. \tag{3.2}$$

In these coordinates the center of mass is

$$\sum_{j=0}^{n} m_j q_j = n z_0 \tag{3.3}$$

so setting $z_0=0$ fixes the center of mass at the origin. The moment of inertia is

$$\begin{aligned} I &= \frac{1}{2}\sum_{j=0}^{n-1} \|q_j\|^2 + \frac{m}{2}\|q_n\|^2 \\ &= \frac{1}{2}\left(n + z_1 + \bar{z}_1 + \sum_{j=1}^{n} z_j \bar{z}_j + \frac{n}{m}\|z_n\|^2 \right) \end{aligned} \tag{3.4}$$

so that the first approximation, the manifold $I = I_0 = n/2$, is given by $z_1 + \bar{z}_1 = 2 \operatorname{Re} z_1 = 0$. Requiring z_1 to be real, $\operatorname{Im} z_1 = 0$, we select a representative from the rotational equivalence class. Thus to the first approximation local coordinates on $\mathcal{S}$ near $[\Omega]$ are $z_0 = z_1 = 0$ and $z_2, z_3, \cdots, z_n$ arbitrary.

Henceforth, set $z_0=0$ and $\operatorname{Im} z_1 = 0$ and let $\operatorname{Re} z_1 = x_1$. From (3.4) we see that $\partial I/\partial x_1(\Omega) = 1 \neq 0$, so by the implicit function theorem we can solve $I = I_0$ for x_1, as a function of the remaining variables. Let $x_1 = \phi(z_2, z_3, \cdots, z_n)$ be this solution. Changing variables by $x_1' = x_1 - \phi(z_2, \cdots, z_n)$, $z_2' = z_2, \cdots, z_n' = z_n$ brings the manifold $I = I_0$ to the hyperplane $x_1 = 0$ locally. Thus $z_2', \cdots, z_n'$ are valid local coordinates on $\mathcal{S}$ near $[\Omega]$.

Computationally we effect this change of variables by using the method of Lie transforms as given by Deprit (1969). We construct the change of variables from the unprimed to the primed variables order by order using the standard normalization procedure. That is, we eliminate the x_1 dependence in I order by order. Henceforth, we will assume that this initial normalization has been carried out, we will ignore z_0 and z_1, and we will drop the primes on the variables.

The next step is to look at the Hessian of the function $\mathcal{U}$ at $[\Omega]$. We can consider (1) as a change to the $(z, \bar{z})$ coordinates or follow Palmore and use the real and imaginary parts of z. We choose the latter for exposition purposes.

Let $z_j = x_j + iy_j$, $\xi = (x_2, \cdots, x_n)^T$, $\eta = (y_2, \cdots, y_n)^T$, $u = (x_2, \cdots, y_n)^T$, and $A = \partial^2 \mathcal{U}/\partial u^2[\Omega]$. Palmore (1976) shows that the Hessian, A, has the relatively simple form

$$A = \begin{pmatrix} B+C & 0 \\ 0 & B-C \end{pmatrix} \tag{3.5}$$

where B and C are $(n-1)\times(n-1)$ matrices, B is a standard diagonal matrix, and C has nonzero entries only on the cross diagonal running northeast. These nonzero entries are given in Appendix A for reference. In Appendix A we give the general formulas for all potentials which vary inversely with the distance of the δ power, so the N-body problem is when $\delta = 1$, and the Kirchhoff problem is the limiting case when $\delta = 0$.

Let $D(n, k)$ (respectively, $D^-(n, k)$), $2 \leqq k \leqq n/2$ be the 2×2 submatrix of $B+C$ (respectively, $B-C$) formed by taking the (k, k), $(k, n+2-k)$, $(n+2-k, k)$, and $(n+2-k, n+2-k)$ entries. In the case that n is even the two diagonals B and C cross in a single entry at the $(n/2+1, n/2+1)$ position; let $D(n, n/2+1)$ (respectively, $D^-(n, n/2+1)$) be the corresponding 1×1 matrix or number. This is a special case, which requires special treatment, and we will typically discuss this case last.

The r.p.r.e. is degenerate when A is singular, which happens when one of the submatrices $D(n, k)$ is singular. Except for the last row of $B \pm C$ all the nonzero entries of $B \pm C$ are linear in m and in fact the determinants of the $D(n, k)$, $2 < k \leqq n/2+1$, are linear in m also. Referring to Appendix A shows that the last row is slightly more complicated, the determinant $D(n, 2)$ has an extraneous factor of $(m+n)$ and another linear factor in m. Thus, there is a unique $m = m(n, k)$, which makes the submatrices $D(n, k)$ and $D^-(n, k)$ singular. In the special case when n is even and $k = n/2+1$ the 1×1 matrix or number $D^-(n, k)$ does not contain m, and so when $m = m(n, k)$ we have $D(n, k) = 0$ but $D^-(n, k) \neq 0$. In this special case the dimension of the kernel of A is one. Let $d(n, k) = \det D(n, k)$ for $m = 0$, $2 < k \leqq n/2+1$. Appendix A also contains the general formulas for $m(n, k)$ and Appendix B contains a table of $m(n, k)$ and $d(n, k)$ for all $3 \leqq n \leqq 12$ for both the N-body problem and the Kirchhoff problem. Recall that $d(n, 2)$ is not defined. The tables in Appendix B are easily generated from the formulas in Appendix A.

Palmore (1973) considered this one-parameter family of r.e. for the $N = n-1$ body problem for $n = 3$, 4 and showed that there was a unique positive value of the mass that makes this r.e. degenerate. In Meyer and Schmidt (1987), we verify this fact and show that additional families of r.e. bifurcate from the original family. Palmore (1976) makes a similar statement about the existence of a unique positive critical mass for all n. From the table in Appendix B, we see that $m(6, 2) \simeq 20.91$ and $m(6, 4) \simeq .00598$, and so this is not the case for $n = 6$. We computed this table all the way up to $n = 20$ and found that as n increases, more and more positive critical masses appear. Moreover, the critical mass of Palmore is $m(n, 2)$ in our notation, and it becomes negative at $n = 7$ and remains negative up to $n = 20$. Later we will show that these positive critical masses give rise to new families of r.e. which bifurcate from the r.p.r.e.

Palmore (1982) also states that there is a unique positive circulation which makes the Kirchhoff potential degenerate for all $n \geqq 3$. Appendix B shows that the uniqueness is false for $8 \leqq n \leqq 12$ and we extended the table to $n = 20$ to find more and more positive critical circulations as n increases. For the Kirchhoff problem the exact formula for the critical circulation $m(n, k)$ takes on a simple form as shown in Appendix A. From this we see that $m(7, 4) \equiv 0$, so one critical circulation is zero. Negative values of the circulation are meaningful and so we investigate these bifurcations in the next section also.

There are several other errors in Palmore (1976), (1982). He also states that the r.e. when $m = 0$ is a nondegenerate maximum of the potential for both problems and for all n. Since we work with the self-potential or the negative of the potential, this would mean that the matrices obtain by deleting the first and last rows and columns of $B \pm C$ are positive definite and in particular the $d(n, k) > 0$ for $2 < k \leqq n/2+1$. This is false when $6 \leqq n \leqq 12$, which can be seen easily by looking at determinants $d(n, k)$ given in Appendix B—in particular $d(6, 4) \simeq -0.036 < 0$. Again we extended this all the way to $n = 20$. We give a simple analytic argument in Appendix D which shows that the potential does not have a minimum at this r.e. when $n = 6$. As noted above the Kirchhoff problem is degenerate when $n = 7$ since $m(7, 4) \equiv 0$. The source of all these errors seems to be in the analysis of the 2×2 submatrices $D(n, k)$.

4. The splitting lemma, reflections, and Hopf's method. There is a simple argument due to Hopf (1942) that establishes a bifurcation without a knowledge of higher-order terms. The analysis of the Hessian given in the previous section along with the symmetry of the potential function is enough to adapt Hopf's argument to the present situation. We present this argument before the discussion of the full normalization to emphasize how little computation is necessary to establish some information about the nature of the bifurcation.

Fix n and k, let $\mu = m - m(n, k)$ and $h = 2n-2$. The special case when n is even and $k = n/2+1$ will be treated at the end, so for now assume we are not in this case. By the analysis of the previous section and the splitting lemma as found in Poston and Stewart (1978) there is a coordinate system η near $[\Omega]$ so that

$$\mathcal{U} = \pm\eta_3^2 \pm \eta_4^2 \pm \cdots \pm \eta_h^2 + G(\eta_1, \eta_2, \mu). \tag{4.1}$$

In catastrophe theory the Lyapunov-Schmidt method is called the splitting lemma. In the next section we discuss in detail how the quadratic terms are brought into the above form and how the function G is computed order by order using Deprit's method of Lie transforms and the second author's algebraic processor POLYPAK.

From the form of the Hessian A in (3.5) and the fact that the submatrices $D(n, k)$ and $D^-(n, k)$ have the same determinant which is linear in the mass m, we see that the quadratic terms of G have the form $\alpha\mu(\eta_1^2+\eta_2^2)/2$ where α is a nonzero constant.

Also $\mathcal{U}$ is invariant under a reflection $\mathcal{R}$ which leaves the regular polygon relative equilibrium fixed. In the original coordinates the reflection is

$$\mathcal{R}: q_j \to \bar{q}_{n-j}, \quad 0 \leqq j < n, \qquad q_n \to \bar{q}_n. \tag{4.2}$$

At one of the critical masses a perturbation in the direction of the kernel of the Hessian is of the form

$$\begin{aligned} q_j &= \omega^j + \omega^{jk} z_k + \omega^{jl} z_l, \qquad k+l = n+2, \\ q_j &= \omega^j + \omega^{jk} z, \qquad k = l = n/2+1. \end{aligned} \tag{4.3}$$

In the first case the z_k and z_l are not independent but are linearly related (essentially conjugates), so one can be used as a coordinate of the perturbation. In the second case the z is arbitrary. Thus we can use z_k or z as a coordinate in the kernel of the Hessian. The action of $\mathcal{R}$ on this subspace is

$$\begin{aligned} &\mathcal{R}: \omega^j + \omega^{jk} z_k + \omega^{jl} z_l \to \omega^j + \omega^{jk} \bar{z}_k + \omega^{jl} \bar{z}_l, \qquad k \neq l, \\ &\mathcal{R}: \omega^j + \omega^{jk} z \to \omega^j + \omega^{jk} \bar{z}, \qquad k = l = n/2+1. \end{aligned} \tag{4.4}$$

Thus in coordinates $\mathcal{R}: z_k \to \bar{z}_k$ or $\mathcal{R}: z \to \bar{z}$, so $\mathcal{R}$ is a reflection on this subspace also. Therefore, we can choose the coordinates η_1 and η_2 so that

$$G(\eta_1, \eta_2, \mu) \equiv G(\eta_1, -\eta_2, \mu). \tag{4.5}$$

This is essentially the same as Lyapunov-Schmidt reduction in the presence of symmetry discussed in Proposition 3.3 of Golubitsky and Schaeffer (1985).

Thus, if $\partial G/\partial\eta_1(\bar{\eta}_1, 0, 0) = 0$, then $\eta = (\bar{\eta}_1, 0, \cdots, 0)$ is a critical point of $\mathcal{U}$. Since $\partial G/\partial\eta_1(0, 0, \mu) = 0$, η_1 is a factor of $\partial G/\partial\eta_1(\eta_1, 0, \mu)$, and so we must solve

$$\begin{aligned} \frac{\partial G}{\partial \eta_1}(\eta_1, 0, \mu) &= \mu\alpha\eta_1 + \eta_1 g(\eta_1, \mu) \\ &= \eta_1(\alpha\mu + g(\eta_1, \mu)) \end{aligned} \tag{4.6}$$

or

$$\alpha\mu + g(\eta_1, \mu) = 0 \tag{4.7}$$

where $g(0, \mu) \equiv 0$. Since $\alpha \neq 0$, the implicit function theorem gives a solution of (3) of the form $\mu = v(\eta_1)$ and so $\eta = (\eta_1, 0, \cdots, 0)$ is a critical point of $\mathcal{U}$ when $\mu = v(\eta_1)$. So we have shown that locally the critical point set of $\mathcal{U}$ in $R^h \times R^1$ consists of two intersecting curves namely $(\eta, \mu) \equiv (0, \mu)$ and $(\eta, \mu) = ((\eta_1, 0, \cdots, 0), v(\eta_1))$. These solutions are symmetric with respect to the reflection $\mathcal{R}$ and are unique in this class. Of course there may be more nonsymmetric solutions.

Hopf's argument just given depends only on the analysis of the Hessian and the symmetry of the system and so is quite easy to apply. The values of $m(n, k)$ and the corresponding α's are easy to compute from the formulas in Appendix A for both the N-body problem and the Kirchhoff problem. Appendix B contains a table of $m(n, k)$ and Appendix C a table of α for $3 \leqq n \leqq 12$. Since the computed values of the α's are nonzero the above result holds in all these cases.

However, this result is rather weak. First of all G could be identically equal to zero, in which case the function $v(\eta_1)$ would be identically zero also. Most people would not call this a bifurcation. The result does not tell how many r.e. are found since the method only looks for symmetric solutions. To overcome the first weakness only a little more computation needs to be carried out.

The full normalization of $\mathcal{U}$ was carried out by the method of Lie transforms using the second author's algebraic processor POLYPAK in almost all cases. The size of the problem grows rapidly with n since (1) the number of variables increases, (2) the number of critical masses increases, and (3) the order to which we must carry out the normalization increases. The first two cause linear growth in complexity whereas the third causes exponential growth in complexity. The full normalization is discussed in the next section.

If we are content to seek only solutions that are symmetric with respect to the x-axis we need only compute the first nonzero term in $G(\eta_1, 0, 0)$ to determine the general nature of the bifurcation. Thus the quest for symmetric solutions grows like a polynomial in n in the generic case.

Using the previous notation as found in (6) assume that

$$g(\eta_1, 0) = -\beta\eta_1^\rho + \cdots \tag{4.8}$$

where $\beta \neq 0$. Then the solution $\mu = v(\eta_1)$ is a solution of

$$\alpha\mu = g(\eta_1, \mu) = \alpha\mu - \beta\eta_1^\rho + \cdots = 0, \tag{4.9}$$

$$\mu = v(\eta_1) = \frac{\beta}{\alpha}\eta_1^\rho + \cdots. \tag{4.10}$$

Now we can decide how many symmetric relative equilibria bifurcate from the regular polygon relative equilibria as μ varies, since we can solve (4.6) for η_1 to find

$$\eta_1 = \sqrt[\rho]{\alpha\mu/\beta + \cdots}. \tag{4.11}$$

Here we use the standard convention about the ρth roots. In particular, if ρ is even, there are two r.e. that bifurcate from the r.p.r.e. for $\mu > 0$ when $\alpha\beta > 0$ or for $\mu < 0$ when $\alpha\beta < 0$. If ρ is odd one r.e. bifurcates from the r.p.r.e. for $\mu < 0$ and one for $\mu > 0$.

In the special case when n is even and $k = n/2 + 1$ the kernel of A has dimension 1 and so the splitting lemma says there are coordinates such that

$$\mathcal{U} = \pm\eta_2^2 \pm \eta_3^2 \pm \cdots \pm \eta_h^2 + G(\eta_1, \mu). \tag{4.12}$$

In the previous case we used the symmetry $\mathcal{R}$ to reduce the problem to that of solving (4.6). In this case we need only look at $\partial G/\partial \eta_1(\eta_1, \mu) = 0$ and proceed exactly as above.

Generically we would expect $\rho = 1$ unless the problem had a further symmetry in which case we would expect $\rho = 2$. We explain this difference in the next section. Thus in the generic case we do not have to compute the function G to high order. Appendix C contains a table of α, β, and ρ for both the N-body and the Kirchhoff problems for $3 \leqq n \leqq 12$, $2 \leqq k \leqq n/2+1$. Note that several entries are missing from the table for the N-body problem since these correspond to negative mass. The N-body problem behaves in a generic manner with ρ being 1 or 2, but the Kirchhoff problem is somewhat unpredictable. Note in Appendix C, when $n = 11$, $k = 6$ that $\rho = 2$ for the N-body problem whereas $\rho = 4$ for the Kirchhoff problem. The Kirchhoff problem is even more degenerate when $n = 4$, $k = 3$. In this case $\rho = 1$ and with it $g(\eta_1, \mu) = 8\mu(\eta_1 + \eta_1^3 + \cdots)$. All the terms have a factor μ and therefore the r.e. exists for $\mu = 0$ with η_1 arbitrary. We will come back to this case in the next section.

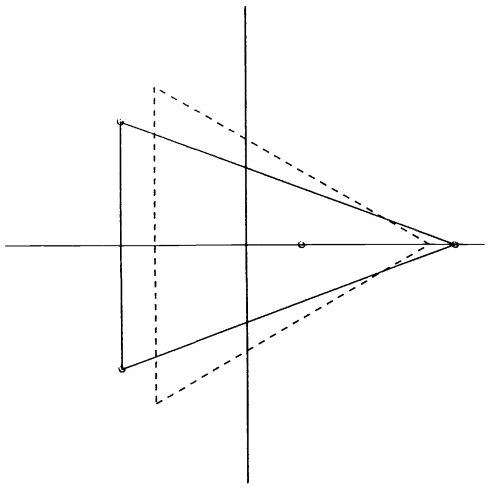


FIG. 1(a). $n = 3$, $k = 2$.

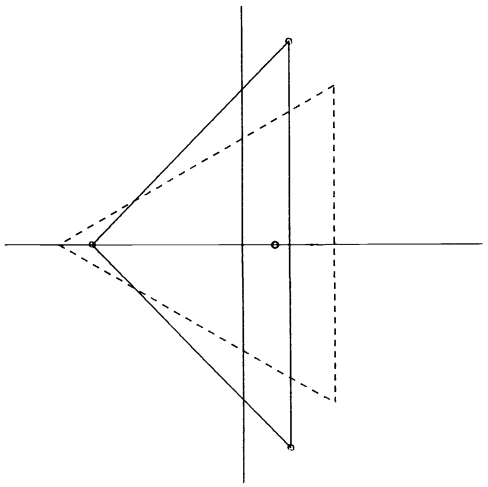


FIG. 1(b). $n = 3$, $k = 2$.

Figure 1 shows the r.e. which bifurcate from the equilateral triangle family. This is the case when $n = 3$, $k = 2$, and $\rho = 1$, so the two isosceles triangle r.e. exist on either side of the critical mass $m(3, 2) \simeq 0.77$ for the 4-body problem or $m(3, 2) = 1$ for the Kirchhoff problem. The acute triangle exists for $m < m(3, 2)$ and the obtuse for $m > m(3, 2)$. Figure 2 shows the r.e. which bifurcate from the square family when $n = 4$, $k = 3$, and $\rho = 2$. Only the kite r.e. shown in Figure 2(a) is symmetric with respect to the x-axis and is established by the above argument. It exists for $m > m(4, 3)$. Figure 3 shows all the r.e. that bifurcate from the duodecigon family when $n = 12$ and for various k and ρ. Only those shown in Figs. 3(a), 3(c), 3(e), 3(g), 3(i), 3(k) (every other one) are symmetric with respect to the x-axis or with respect to $\mathcal{R}$. Figure 3 shows the special case when n is even and $k = n/2+1$. These are the ones established by this argument. Note that all of the r.e. in Fig. 3 have an axis of symmetry even though in some cases it is difficult to see at first glance. We will discuss these figures more in the next section.

5. Symmetries and higher-order normalization. In the special case when n is even and $k = n/2+1$ the Hessian A does not have a two-dimensional kernel, and so the

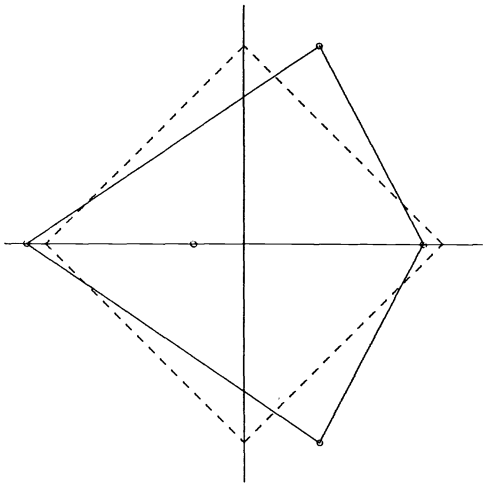


FIG. 2(a). $n=4$, $k=2$.

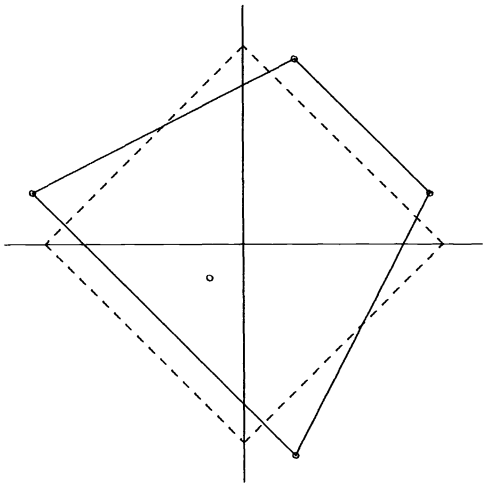


FIG. 2(b). $n=4$, $k=2$.

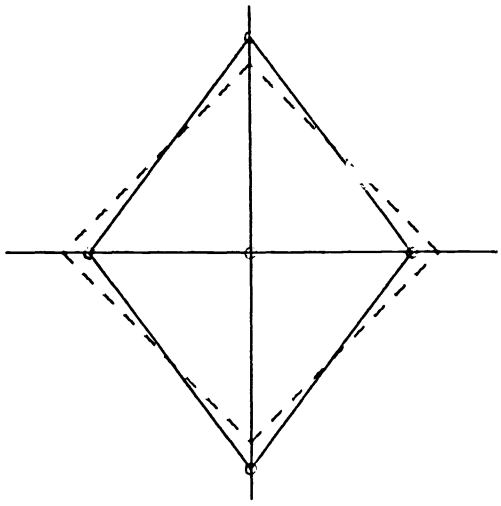


FIG. 2(c). $n=4$, $k=3$.

discussion of the previous section is complete for this case. Thus we will assume that $k \neq n/2+1$ throughout this section.

Let $\xi=(\xi_1, \cdots, \xi_h)^T=(x_2, \cdots, x_n, y_2, \cdots, y_n)^T$, where $h=2n-2$, are the Palmore coordinates discussed in § 2. As before fix n and k and let $\mu=m-m(n,k)$. Obviously $\mathcal{U}$ is invariant under the symmetries of the regular polygon with n sides; that is, there is a subgroup D_n of the orthogonal group $O(h, R)$ which is isomorphic to the dihedral group such that

$$\mathcal{U}(D\xi, \mu)=\mathcal{U}(\xi, \mu) \tag{5.1}$$

for all $D \in D_n$ and all small ξ and μ.

When $\mu=0$, the Hessian A of $\mathcal{U}$ at $\xi=0$ has a two-dimensional kernel and therefore there is an orthogonal matrix O, such that $O^TAO=\operatorname{diag}(0, 0, \lambda_3, \cdots, \lambda_h)$ where $\lambda_i \neq 0$ for $3 \leqq i \leqq h$. Let $O_2=\operatorname{diag}(1, 1, 1/\sqrt{|\lambda_3|}, \cdots, 1/\sqrt{|\lambda_h|})$ so $B=O^TAO=$

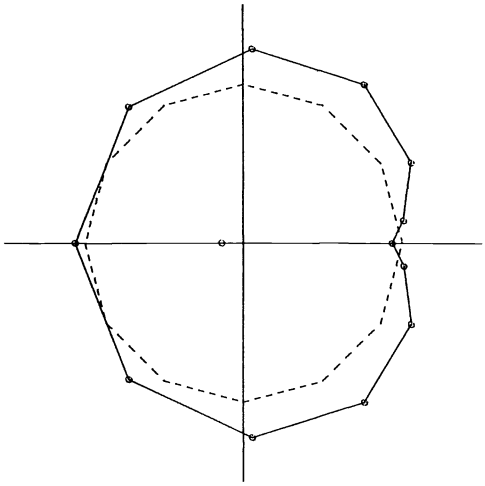


FIG. 3(a). $n = 12$, $k = 2$.

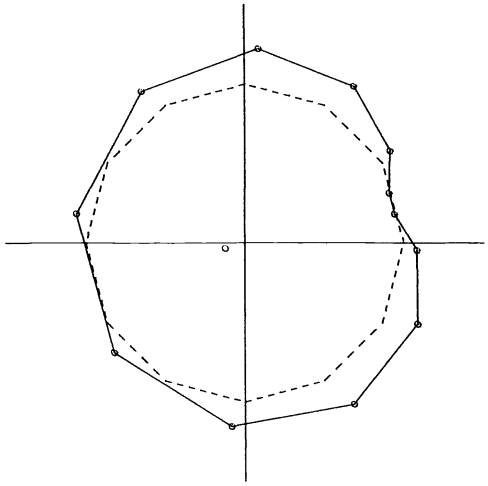


FIG. 3(b). $n = 12$, $k = 2$.

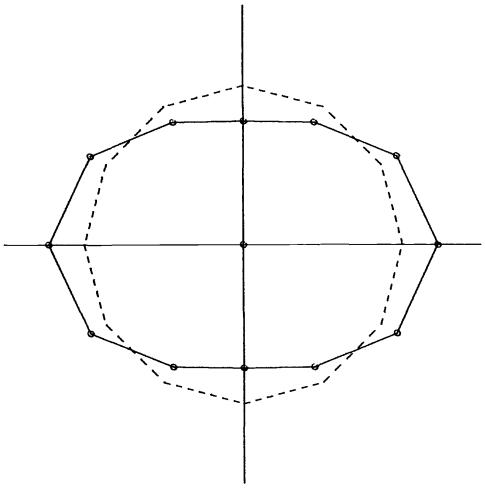


FIG. 3(c). $n = 12$, $k = 3$.

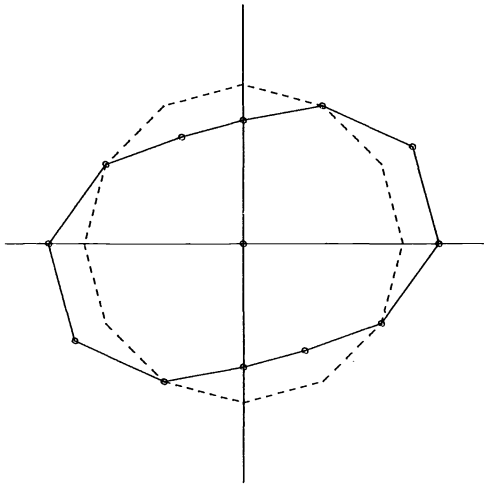


FIG. 3(d). $n = 12$, $k = 3$.

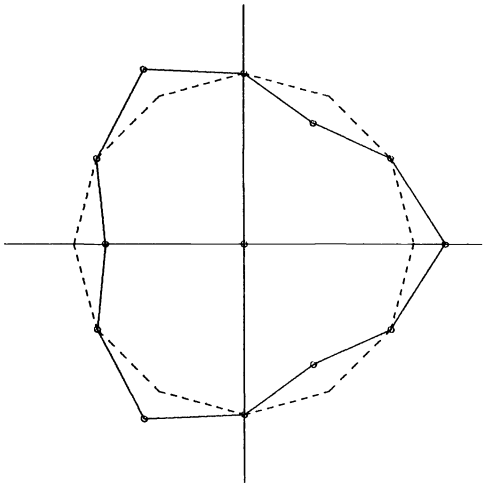


FIG. 3(e). $n = 12$, $k = 4$.

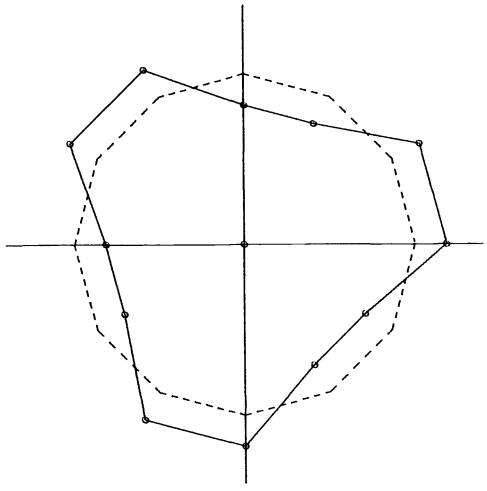


FIG. 3(f). $n = 12$, $k = 4$.

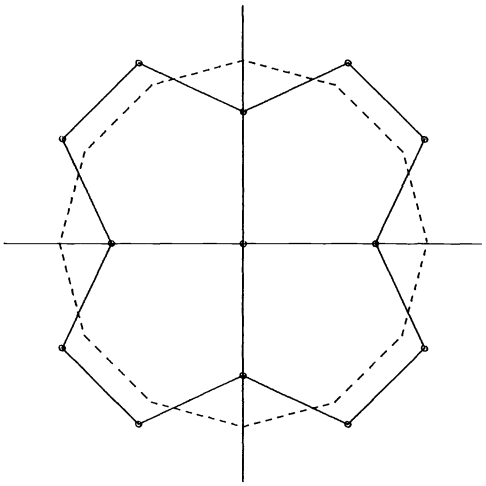


FIG. 3(g). $n = 12$, $k = 5$.

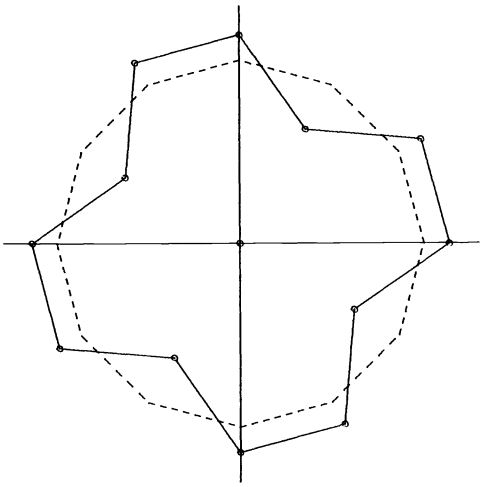


FIG. 3(h). $n = 12$, $k = 5$.

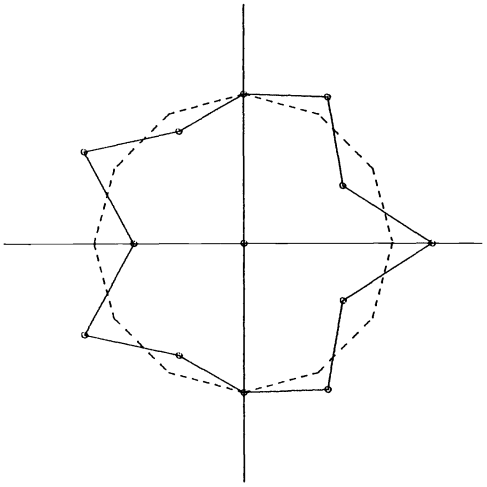


FIG. 3(i). $n = 12$, $k = 6$.

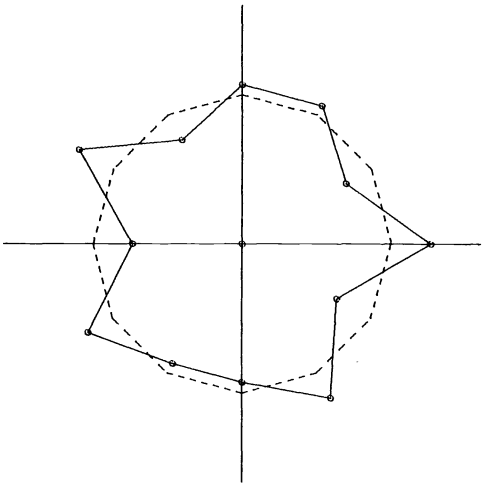


FIG. 3(j). $n = 12$, $k = 6$.

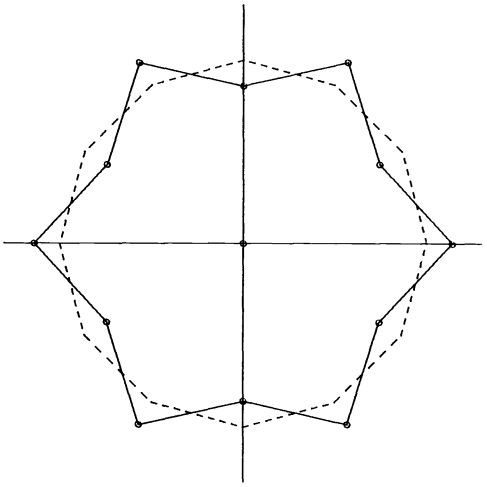


FIG. 3(k). $n = 12$, $k = 7$.

diag $(0, 0, \pm 1, \cdots, \pm 1)$ where $O = O_1 O_2$. If we change coordinates by $\xi = O\zeta$ then the Hessian of $\mathcal{U}$ in these coordinates is B or the quadratic part of $\mathcal{U}$ is as in (4.1). Usually, we use the same symbol for a function in different coordinates, but for the moment let $\mathcal{U}'(\zeta, \mu) = \mathcal{U}(O\zeta, \mu)$ so

$$\begin{aligned} &\mathcal{U}(D\xi, \mu) = \mathcal{U}(\xi, \mu), \\ &\mathcal{U}(OO^{-1}DO\zeta, \mu) = \mathcal{U}(O\zeta, \mu), \\ &\mathcal{U}'(O^{-1}DO\zeta, \mu) = \mathcal{U}'(\zeta, \mu), \end{aligned} \tag{5.2}$$

or

$$\mathcal{U}'(D'\zeta, \mu) = \mathcal{U}'(\zeta, \mu) \tag{5.3}$$

where $D' \in \mathcal{D}_n' = O^{-1}\mathcal{D}_n O$. From (5.3) D' leaves the Hessian of $\mathcal{U}'$ invariant and so $D'^T B D' = B$. This and the special form of O implies D' is of the form

$$D' = \begin{pmatrix} E & 0 \\ 0 & F \end{pmatrix} \tag{5.4}$$

where E is a 2×2 orthogonal matrix and F is some nonsingular $(h-2) \times (h-2)$ matrix. The set of such E's form a subgroup $\mathcal{E}$ of $O(2, R)$ which is clearly isomorphic to the subgroup of $\mathcal{D}_n'$ obtained by letting F be the identity matrix in (5.4). Since $\mathcal{E}$ is isomorphic to a subgroup of a dihedral group and as we saw in the last section contains a reflection, it must be isomorphic to a dihedral group whose order divides $2n$. The order of this group depends on n and k, and the precise dependence will be given at the end of this section along with the discussion of the specific findings.

Let ε be a formal parameter and consider

$$\mathcal{U}_*(\zeta, \mu, \varepsilon) = \sum_{i=0}^{\infty} \left(\frac{\varepsilon^i}{i!}\right) \mathcal{U}_i^0(\zeta, \mu) \tag{5.5}$$

where $\mathcal{U}_*(\zeta, \mu, 1) = \mathcal{U}(\zeta, \mu)$ and $\mathcal{U}_i^0$ is a homogeneous polynomial in ζ of degree $i+2$. The method of Lie transform given by Deprit (1969) constructs a near identity change of variables

$$\zeta = \zeta(\eta, \mu, \varepsilon) = \eta + \cdots$$

where ζ is the general solution of the differential equation

$$\frac{d\zeta}{d\varepsilon} = W(\zeta, \mu, \varepsilon), \qquad \zeta|_{\varepsilon=0} = \eta. \tag{5.6}$$

If the function W has the formal expansion

$$W(\zeta, \mu, \varepsilon) = \sum_{l=0}^{\infty} \left(\frac{\varepsilon^l}{l!}\right) W_{l+1}(\zeta, \mu) \tag{5.7}$$

then in the new coordinates

$$\begin{aligned} \mathcal{U}^*(\eta, \mu, \varepsilon) &= \mathcal{U}_*(\zeta(\eta, \mu, \varepsilon), \mu, \varepsilon) \\ &= \sum_{j=0}^{\infty} \left(\frac{\varepsilon^j}{j!}\right) \mathcal{U}_0^j(\zeta, \mu). \end{aligned} \tag{5.8}$$

The functions $\mathcal{U}_*$ and $\mathcal{U}^*$ are related by the double index array $\{\mathcal{U}_i^j\}$, which agrees with the previous definitions when either i or j is zero and are related by the recursive

relation

$$\mathcal{U}_i^j = \mathcal{U}_{i-1}^{j+1} + \sum_{k=0}^{i} \binom{i}{k} [\mathcal{U}_{i-k}^{j-1}, W_{k+1}], \tag{5.9}$$

where [,] is the Lie derivative operator on functions given by

$$[\mathcal{U}, W] = \frac{\partial \mathcal{U}}{\partial \zeta} W. \tag{5.10}$$

Let $\mathcal{P}_k$ be the space of homogeneous polynomials of degree $k+2$ in $\zeta_1, \cdots, \zeta_h$ with coefficients which are smooth in μ. Let $\mathcal{K}_k$ be the subspace of $\mathcal{P}_k$ of homogeneous polynomials in ζ_1 and ζ_2 only and $\mathcal{R}_k = \zeta_3 \mathcal{P}_{k-1} + \cdots + \zeta_h \mathcal{P}_{k-1}$ so that $\mathcal{P}_k = \mathcal{K}_k \oplus \mathcal{R}_k$. Since $\mathcal{U}_0^0 = \frac{1}{2}\zeta^T B \zeta$ the operator $L: W \to [\mathcal{U}_0^0, W]$ defines a self-map of $\mathcal{P}_k$ with kernel $\mathcal{K}_k$ and range $\mathcal{R}_k$. By a standard argument in normal form theory, we can find a formal series for W so that $\mathcal{U}^*$ is in normal form, i.e., $\mathcal{U}_0^k \in K_k$ for all $k \geqq 1$. That is, the higher-order terms in $\mathcal{U}^*$ depend only on η_1 and η_2. This argument is found in Meyer and Schmidt (1977), for example. This is the formal version of the splitting lemma.

Moreover, the normalizing generating function W satisfies $W_k \in \mathcal{R}_k$ so that the function W is zero on $\Xi = \{\xi: \xi_3 = \cdots = \xi_h = 0\}$. Thus Ξ is an invariant surface for (7) or the change of variables (6) fixes Ξ or $\Xi = \{\eta: \eta_3 = \cdots = \eta_h = 0\}$. This means that the new function $\mathcal{U}^*$ is invariant under the linear action defined by the matrices of the form (5.4) with $F = I$, the identity matrix.

We see that the normal form for $\mathcal{U}$ is the same as given by the splitting lemma in formula (4.1). Moreover, if the normalization is carried out as outlined above, the higher-order terms (i.e., G in 4.1) are invariant under the standard action of $\mathcal{E}$ on the plane.

Let $\mathcal{E}$ have order $2d$ where d divides n. Appendix C has a table giving d for various n and k. Consider the η_1, η_2 plane as the complex plane by setting $w = \eta_1 + i\eta_2$ and let $\phi = \exp(2\pi i/d)$ be a primitive dth root of unity. By the above, we are reduced to studying the critical points of

$$\Gamma(w, \bar{w}, \mu) = G(\eta_1, \eta_2, \mu) \tag{5.11}$$

where Γ is invariant under the action of $\mathcal{E}$ or

$$\Gamma(\phi w, \overline{\phi w}, \mu) = \Gamma(w, \bar{w}, \mu), \qquad \Gamma(w, \bar{w}, \mu) = \Gamma(\bar{w}, w, \mu). \tag{5.12}$$

The only terms in a Taylor expansion which satisfy the conditions in (5.12) are of the form

$$(w\bar{w})^i w^{dj} \quad \text{or} \quad (w\bar{w})^i \bar{w}^{dj} \tag{5.13}$$

where i and j are integers. Thus a typical expansion of Γ would look like this:

$$\Gamma(w, \bar{w}, \mu) = p_1(w\bar{w}) + p_2(w\bar{w})^2 + \cdots + (1/d) q_1(w^d + \bar{w}^d) + \cdots. \tag{5.14}$$

The p's and q's are real functions of μ. By the analysis of the Hessian given in § 3, $p_1(\mu) = a\mu + \cdots$ where a is a nonzero constant. Assume we are in the generic case so that $q_1(0) \neq 0$ and in addition $p_2(0) \neq 0$ when $d > 4$ and $p_2(0) \neq q_1(0)$ when $d = 4$.

Case 1. $d = 3$. Let $q_1(0) = b$, so we must solve

$$\begin{aligned} &\frac{\partial \Gamma}{\partial w} = a\mu\bar{w} + bw^2 + \cdots = 0, \\ &a\mu(w\bar{w}) + bw^3 + \cdots = 0, \\ &a\mu r^2 + br^3 \exp(i3\theta) + \cdots = 0 \end{aligned} \tag{5.15}$$

where $w = r\exp(i\theta)$. Thus, by the implicit function theorem Γ has critical points at $r = \mp a\mu/b + \cdots$ when $\exp(i3\theta) = \pm 1$. That is, three critical points move linearly away from the origin (or the r.p.r.e.) as μ varies from zero. This case occurs when $n = 3$, $k = 2$ (Fig. 1(a) shows a solution where $\exp(i3\theta) = +1$ and Fig. 1(b) where $\exp(i3\theta) = -1$.); when $n = 9$, $k = 4$; $n = 12$, $k = 5$ (Fig. 3(g) shows a solution where $\exp(i3\theta) = +1$ and Fig. 3(h) where $\exp(i3\theta) = -1$) for both problems, and when $n = 6$, $k = 3$ for the Kirchhoff problem. See Appendix C. In the notation of the previous section $\alpha = a$, $\beta = b$, and $\rho = 1$.

Case 2. $d \geqq 5$. Let $p_2(0) = b \neq 0$ and $q_1(0) = c \neq 0$, so we must solve

$$\begin{aligned} \frac{\partial \Gamma}{\partial w} &= a\mu\bar{w} + b(w\bar{w})^2 + \cdots + cw^{d-1} + \cdots = 0, \\ a\mu r^2 + br^4 + \cdots + cr^d \exp(id\theta) + \cdots &= 0. \end{aligned} \tag{5.16}$$

By the implicit function theorem Γ has critical points at

$$r = \sqrt{-a\mu/b + \cdots}, \qquad \exp(id\theta) = \pm 1. \tag{5.17}$$

That is, Γ has $2d$ nonzero critical points for $\mu > 0$ and none for $\mu < 0$ when $ab < 0$ and vice versa when $ab > 0$. These solutions fall into two families of d each depending on the sign of $\exp(id\theta)$. The families move away from the origin like the square root of μ. For most n and k, we have $d \geqq 5$ (see Appendix C). For $n = 12$, Figs. 3(a), 3(i) show the solutions where $\exp(i12\theta) = +1$ and Figs. 3(b), 3(j) show the solutions where $\exp(i12\theta) = -1$. In the notation of the previous section $\alpha = a$, $\beta = b$, and $\rho = 2$.

Case 3. $d = 4$. Using the above notation, we must solve

$$a\mu r^2 + (b + c\exp(i4\theta))r^4 + \cdots = 0 \tag{5.18}$$

and so there are solutions of the form

$$r = \sqrt{-a\mu/(b \pm c) + \cdots}, \qquad \exp(i4\theta) = \pm 1. \tag{5.19}$$

If $b \pm c$ are of one sign then there are eight solutions for μ on one side of zero as in Case 2. This happens when $n = 12$, $k = 4$. Figure 3(e) shows a solution when $\exp(i4\theta) = +1$ and Fig. 3(f) shows a solution where $\exp(i4\theta) = -1$. If $b \pm c$ have two signs then there are four solutions when μ is negative and when μ is positive as in Case 1. This happens when $n = 4$, $k = 2$. Figure 2(a) shows a solution when $\exp(i4\theta) = +1$ and Fig. 2(b) shows a solution when $\exp(i4\theta) = -1$.

To understand the relationship between n, k, and the order d of the rotational subgroup which acts on the two-dimensional subspace, we proceed as we did in the previous section when we discussed the reflection symmetry. $\mathcal{U}$ is invariant also under a rotation $\mathcal{C}$ which leaves the regular polygon relative equilibrium fixed. In the original coordinates the rotation is

$$\mathcal{C}: q_j \to \omega q_{j-1}, \quad 0 \leqq j < n, \qquad \mathcal{C}: q_n \to q_n. \tag{5.20}$$

This rotation $\mathcal{C}$ with the reflection $\mathcal{R}$ generates the symmetry group of $\mathcal{U}$ which also fixes the r.p.r.e. Ω. At one of the critical masses a perturbation in the direction of the kernel of the Hessian is of the form

$$\begin{aligned} q_j &= \omega^j + \omega^{jk} z_k + \omega^{jl} z_l, \qquad k + l = n + 2, \\ q_j &= \omega^j + \omega^{jk} z, \qquad k = l = n/2 + 1. \end{aligned} \tag{5.21}$$

In the first case the z_k and z_l are not independent but are linearly related (essentially conjugates), so one can be used as a coordinate of the perturbation. In the second case the z is arbitrary. Thus we can use z_k or z as a coordinate in the kernel of the Hessian. The action of $\mathcal{O}$ on this subspace is

$$\begin{aligned} &\mathcal{O}: \omega^j + \omega^{jk} z_k + \omega^{jl} z_l \to \omega^j + \omega^{jk}(\omega^{1-k}) z_k + \omega^{jl}(\omega^{l-1}) z_l, \qquad k \neq l, \\ &\mathcal{O}: \omega^j + \omega^{jk} z \to \omega^j + \omega^{jk}(\omega^{1-k}) z, \qquad k = l = n/2 + 1. \end{aligned} \tag{5.22}$$

Thus in coordinates $\mathcal{O}: z_k \to (\omega^{1-k}) z_k$ or $\mathcal{O}: z \to (\omega^{1-k}) z$, so $\mathcal{O}$ is a rotation on this subspace also; but it does not necessarily generate the full symmetry group. The order of the rotation group generated by $\mathcal{O}$ on this subspace is d where $(k-1)d \equiv 0 \bmod n$. Appendix C lists n, k, and d for all cases $3 \leqq n \leqq 12$.

Consider Figs. 3(e) and 3(f) for example where $n = 12$, $k = 4$, and $d = 4$. By rotating these figures by $l2\pi/12$ for $l = 0, 1, \cdots, 11$ we obtain $d = 4$ distinct r.e. which have symmetry given by the dihedral group D_3, because $k - 1 = 3$. Contrast that with Figs. 3(a) and 3(b) where $n = 12$, $k = 2$, $d = 12$. When we rotate these figures by $l2\pi/12$, $l = 0, \cdots, 11$ we obtain $d = 12$ distinct r.e., which have the symmetries of the dihedral group D_1 (generated by a single reflection), since $k - 1 = 1$. The other figures follow the same pattern.

Finally we will consider the degeneracy in the Kirchhoff problem when $n = 4$, $k = 3$. In this case the symmetry with respect to both axes is preserved. Since the moment of inertia has to remain constant, r.e. can only be formed by a rhombus with the fifth vortex at the center. The coordinates of the five vortices are

$$\begin{aligned} q_0 &= -q_2 = 1 + x_0, \\ q_1 &= q_3 = i\sqrt{1 - 2x_0 - x_0^2}, \\ q_4 &= 0. \end{aligned}$$

Let m be the value of the vorticity at the origin, then the potential function (2.8) turns out to be

$$-\tfrac{7}{8}\log 2 - (1+2m)\{\log(1+x_0) + \tfrac{1}{2}\log(1 - 2x_0 - x_0^2)\}.$$

For $m \neq -\frac{1}{2}$ the potential function has extrema at $x_0 = 0$ and $x_0 = -\frac{1}{2}$. For $m = -\frac{1}{2}$ the potential function is independent of x_0 and therefore any rhombus can serve as a r.e. for the Kirchhoff problem. See Fig. 2(c).

Appendix A. Entries in the Hessian.

$$B = \begin{pmatrix} b_2 & & & \\ & b_3 & & \\ & & \ddots & \\ & & & b_n \end{pmatrix}, \qquad C = C^T = \begin{pmatrix} & & & c_n \\ & & \iddots & \\ & c_3 & & \\ c_2 & & & \end{pmatrix},$$

$$b_k = \frac{n}{2}(-R_k - \gamma m), \qquad k = 2, 3, \cdots, n-1,$$

$$b_n = \frac{n}{2}\frac{m+n}{m}(-\gamma m - \delta n - S),$$

$$c_k = \frac{n}{2}(T_k - \gamma m), \qquad k = 3, 4, \cdots, n-1,$$

$$c_k = -\frac{n}{2}(m+n)\gamma, \qquad k = 2, n$$

where $\gamma = \delta + 2$ and

$$S = \frac{1}{2^\delta} \sum_{r=1}^{n-1} \frac{1}{\sin^\delta (\pi r/n)},$$

$$R_k = \frac{\delta}{2^{\delta+1}} \sum_{r=1}^{n-1} \frac{\sin^2 (\pi rk/n)}{\sin^\gamma (\pi r/n)} + S,$$

$$T_k = \frac{\gamma}{2^{\delta+1}} \sum_{r=1}^{n-1} \frac{\sin (\pi rk/n) \sin (\pi r(k-2)/n)}{\sin^\gamma (\pi r/n)},$$

$$m(n, 2) = \frac{R_k(\delta n + S)}{\gamma(2n - R_2 - S)},$$

$$m(n, k) = \frac{T_k^2 - R_k R_l}{\gamma(R_k + R_l + 2T_k)}, \qquad k + l = n + 2, \quad k \neq l,$$

$$m(n, k) = \frac{T_k - R_k}{2\gamma}, \qquad 2k = n + 2.$$

In the Kirchhoff problem, $\delta = 0$, the formulas for m simplify to

$$m(n, 2) = \tfrac{1}{4}(n-1)^2,$$

$$m(n, k) = \tfrac{1}{4}\{(k-2)(n-k) - n + 1\}, \qquad k = 3, 4, \cdots, (n+2)/2.$$

Appendix B. Critical masses and subdeterminants.

		Kirchhoff		$n+1$ body problem	
n	k	$m(n, k)$	$d(n, k)$	$m(n, k)$	$d(n, k)$
3	2	1.000E+00	—	7.705E−01	—
4	2	2.250E+00	—	2.380E+00	—
4	3	−5.000E−01	2.000E+00	−2.500E−01	1.500E+00
5	2	4.000E+00	—	6.478E+00	—
5	3	−5.000E−01	1.200E+01	−2.442E−01	1.144E+01
6	2	6.250E+00	—	2.091E+01	—
6	3	−5.000E−01	1.600E+01	−2.201E−01	1.577E+01
6	4	−2.500E−01	1.000E+00	5.983E−03	−3.590E−02
7	2	9.000E+00	—	−6.433E+02	—
7	3	−5.000E−01	2.000E+01	−1.814E−01	1.800E+01
7	4	0.0	0.0	3.242E−01	−4.342E+01
8	2	1.225E+01	—	−3.793E+01	—
8	3	−5.000E−01	2.400E+01	−1.306E−01	1.686E+01
8	4	2.500E−01	−1.500E+01	6.980E−01	−1.301E+02
8	5	5.000E−01	−2.000E+00	9.963E−01	−5.978E+00
9	2	1.600E+01	—	−2.544E+01	—
9	3	−5.000E−01	2.800E+01	−6.937E−02	1.116E+01
9	4	5.000E−01	−3.600E+01	1.119E+00	−2.725E+02
9	5	1.000E+00	−8.000E+01	1.774E+00	−5.133E+02
10	2	2.025E+01	—	−2.172E+01	—
10	3	−5.000E−01	3.200E+01	1.064E−03	−2.068E−01
10	4	7.500E−01	−6.300E+01	1.581E+00	−4.819E+02
10	5	1.500E+00	−1.440E+02	2.641E+00	−1.002E+03
10	6	1.750E+00	−7.000E+00	3.012E+00	−1.807E+01
11	2	2.500E+01	—	−2.027E+01	—

Appendix B. Critical masses and subdeterminants (*cont.*)

11	3	−5.000E−01	3.600E+01	7.969E−02	−1.830E+01
11	4	1.000E+00	−9.600E+01	2.080E+00	−7.687E+02
11	5	2.000E+00	−2.240E+02	3.588E+00	−1.708E+03
11	6	2.500E+00	−3.000E+02	4.391E+00	−2.340E+03
12	2	3.025E+01	—	−1.974E+01	—
12	3	−5.000E−01	4.000E+01	1.657E−01	−4.411E+01
12	4	1.250E+00	−1.350E+02	2.611E+00	−1.143E+03
12	5	2.500E+00	−3.200E+02	4.605E+00	−2.665E+03
12	6	3.250E+00	−4.550E+02	5.894E+00	−3.949E+03
12	7	3.500E+00	−1.400E+01	6.338E+00	−3.803E+01

Appendix C. Coefficients.

			Kirchhoff			$n+1$ body problem		
n	k	d	ρ	α	β	ρ	α	β
3	2	3	1	−1.200E+00	1.717E+01	1	−1.921E+00	1.562E+01
4	2	4	2	−7.843E−01	−4.919E+01	2	−9.756E−01	−5.327E+01
4	3	3	1	8.000E+00	0	—	—	—
5	2	5	2	−5.769E−01	5.858E+00	2	−3.773E−01	1.132E+01
5	3	5	2	1.000E+01	−7.500E+00	—	—	—
6	2	6	2	−4.541E−01	5.456E+00	2	−7.906E−02	8.977E+00
6	3	3	1	1.200E+01	−1.697E+01	—	—	—
6	4	4	2	1.200E+01	−4.500E+01	2	1.800E+01	−7.850E+01
7	2	7	2	−3.733E−01	5.312E+00	—	—	—
7	3	7	2	1.400E+01	−2.100E+02	—	—	—
7	4	7	?	1.400E+01	??	2	2.098E+01	−1.169E+02
8	2	8	2	−3.165E−01	5.298E+00	—	—	—
8	3	4	2	1.600E+01	−3.840E+02	—	—	—
8	4	8	2	1.600E+01	9.000E+01	2	2.393E+01	−1.930E+02
8	5	4	2	1.600E+01	−2.560E+02	2	2.400E+01	−5.497E+02
9	2	9	2	−2.744E−01	5.360E+00	—	—	—
9	3	9	2	1.800E+01	−2.100E+02	—	—	—
9	4	3	1	1.800E+01	−5.728E+01	1	2.688E+01	−6.443E+01
9	5	9	2	1.800E+01	−2.550E+02	2	2.699E+01	−6.070E+02
10	2	10	2	−2.420E−01	5.469E+00	—	—	—
10	3	5	2	2.000E+01	−2.400E+02	2	2.930E+01	−2.550E+02
10	4	10	2	2.000E+01	−1.102E+03	2	2.982E+01	−1.325E+03
10	5	10	2	2.000E+01	−4.800E+02	2	2.997E+01	−1.685E+03
10	6	4	2	2.000E+01	−8.750E+02	2	3.000E+01	−2.253E+03
11	2	11	2	−2.164E−01	5.609E+00	—	—	—
11	3	11	2	2.200E+01	−2.772E+02	2	3.217E+01	−3.099E+02
11	4	11	2	2.200E+01	−9.900E+02	2	3.276E+01	−1.479E+03
11	5	11	4	2.200E+01	3.831E+04	2	3.294E+01	−2.005E+03
11	6	11	2	2.200E+01	−7.425E+02	2	3.299E+01	−2.312E+03
12	2	12	2	−1.956E−01	5.772E+00	—	—	—
12	3	6	2	2.400E+01	−3.200E+02	2	3.504E+01	−3.739E+02
12	4	4	2	2.400E+01	−1.836E+03	2	3.570E+01	−2.628E+03
12	5	3	1	2.400E+01	−1.357E+02	1	3.591E+01	−1.759E+02
12	6	12	2	2.400E+01	−1.925E+03	2	3.598E+01	−6.147E+03
12	7	4	2	2.400E+01	−2.304E+03	2	3.600E+01	−6.947E+03

Appendix D. The potential of the hexagon configuration. Consider the one-parameter perturbation of the hexagon configuration (with no particle at the centroid) in the n-body problem given by

$$q_0=(1+\varepsilon)(2,0), \qquad q_1=\sqrt{(1-2\varepsilon-\varepsilon^2)}(1,\sqrt{3}),$$
$$q_2=(1+\varepsilon)(-1,\sqrt{3}), \qquad q_3=\sqrt{(1-2\varepsilon-\varepsilon^2)}(-2,0),$$
$$q_4=(1+\varepsilon)(-1,-\sqrt{3}), \qquad q_5=\sqrt{(1-2\varepsilon-\varepsilon^2)}(1,-\sqrt{3}).$$

This perturbation has been chosen to keep the moment of inertia, I, fixed. From the symmetry,

$$\begin{aligned} U &= \frac{6}{\|q_0-q_1\|}+\frac{3}{\|q_0-q_2\|}+\frac{3}{\|q_1-q_3\|}+\frac{3}{\|q_0-q_3\|} \\ &= 3\{1-\varepsilon^2+\cdots\}+\frac{\sqrt{3}}{2}\{1-\varepsilon+\varepsilon^2+\cdots\}+\frac{\sqrt{3}}{2}\{1+\varepsilon+2\varepsilon^2+\cdots\}+\frac{3}{4}\left\{1+\frac{1}{2}\varepsilon^2+\cdots\right\} \\ &= \left(\frac{15}{4}+\sqrt{3}\right)-\frac{3}{8}(7-4\sqrt{3})\varepsilon^2+\cdots \\ &\simeq 5.48-0.0269\varepsilon^2+\cdots. \end{aligned}$$

Thus U initially decreases along this family and so the hexagon configuration is not a minimum of the (self-) potential.

REFERENCES

A. DEPRIT (1969), *Canonical transformation depending on a small parameter*, Celestial Mech., 72, pp. 173-179.

O. DZIOBEK (1900), *Über einen merkwürdigen Fall des Vielkörperproblem*, Astronom. Nachr., 152, pp. 32-46.

L. EULER (1767), *De motu restilineo trium corporum se mutus attrahentium*, Novi Comm. Acad. Sci. Imp. Petrop., 11, pp. 144-151.

M. GOLUBITSKY AND D. G. SCHAEFFER (1985), *Singularities and Groups in Bifurcation Theory*, Springer-Verlag, New York.

E. HOPF (1942), *Abzweigung einer periodschen Lösung von einer stationären Lösung eines Differential System*, Ber. Verh. Sachs. Acad. Wiss. Leipzig Math.-Nat. Kl., 95(1), pp. 3-22.

R. HOPPE (1879), *Erweiterung der bekannten Speziallösung des Dreikörper Problems*, Archiv Math. und Phys., 64, pp. 218-223.

G. KIRCHHOFF (1897), *Vorlesungen über Mechanik*, Druck und Verlag von B. G. Teubner, Leipzig.

J. L. LAGRANGE (1772), *Essai sur le problem des trois corps*, Oeuvers, 6, pp. 272-292.

R. LEHMANN-FILHES (1891), *Über zwei Fälle des Vielkörpersproblem*, Astronom. Nachr., 127, pp. 137-144.

K. R. MEYER AND D. S. SCHMIDT (1977), *Entrainment domains*, Funkcial. Ekvac., 20, pp. 171-192.

——— (1988), *Bifurcations of relative equilibria in the 4 and 5 body problems*, to appear in Ergodic Theory, Dynamical Systems.

R. MOECKEL (1985), *Relative equilibria of the four-body problem*, Ergodic Theory Dynamical Systems, 5, pp. 417-435.

F. R. MOULTON (1910), *The straight line solutions of the problem of n-bodies*, Bull. Amer. Math. Soc., 13, pp. 324-335.

J. I. PALMORE (1973), *Relative equilibria of the n-body problem*, Ph.D. thesis, University of California, Berkeley, CA.

——— (1975), *Classifying relative equilibria* II, Bull. Amer. Math. Soc., 81, pp. 489-491.

——— (1976), *Measure of degenerate relative equilibria* I, Ann. of Math., 140, pp. 421-429.

——— (1982), *Relative equilibria of vortices in two dimensions*, Proc. Nat. Acad. Sci. USA, 79, pp. 716-718.

T. POSTON AND I. STEWART (1978), *Catastrophe Theory and its Applications*, Pitman, Boston.

C. SIMO (1977), *Relative equilibrium solutions in the four body problem*, Celestial Mech., 18, pp. 165-184.

S. SMALE (1970), *Topology and mechanics.* II, *The planar n-body problem*, Invent. Math., 11, pp. 45-64.

D. SAARI (1980), *On the pole and properties of n-body central configurations*, Celestial Mech., 21, pp. 9-20.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

004

AN EXISTENCE RESULT FOR THE ELECTROPAINTING PROBLEM*

PIERLUIGI COLLI† AND LUC OSWALD‡

Abstract. A time-dependent family of harmonic problems with boundary condition $h(\partial\varphi/\partial\nu)=\varphi$, where h is a function dependent on the history of φ, models an electropaint process. It is proven that the problem has a weak solution $\{\varphi(x,t), h(x,t)\}$ and $\lim_{t\to\infty}\varphi(x,t)$ exists and coincides with the solution of an appropriate Signorini problem.

Key words. electropainting process, time-dependent family of harmonic problems

AMS (MOS) subject classifications. 35D05, 35J65, 35R35

Introduction. In this paper we are interested in an electropaint process model, which was introduced by Aitchison, Lacey, and Shillor [ALS] as an approximation of the physical problem. Electropainting is a commonly used method for painting metal surfaces: a workpiece is immersed in an electrolyte solution and a potential difference is applied between the workpiece and the outside boundary of the bath. This induces a deposition of ions on the surface of the metal, which is subsequently painted (for more details, see [ALS]).

Let us describe the mathematical problem. We denote by Γ the surface of the workpiece and by S the boundary of the bath: Ω will be the region occupied by the electrolyte solution. Then $\partial\Omega=S\cup\Gamma$ and Ω is a domain in $\mathbb{R}^N$, $N\geqq 2$. Let φ be the electric potential in Ω and h the thickness of the paint layer on Γ. The problem is to find $\varphi(x,t)$, $h(x,t)$ such that

$$\Delta\varphi=0 \quad \text{in } \Omega, \qquad t>0, \tag{0.1}$$

$$\varphi=1 \quad \text{on } S, \qquad t>0, \tag{0.2}$$

$$h\frac{\partial\varphi}{\partial\nu}=\varphi \quad \text{on } \Gamma, \qquad t>0, \tag{0.3}$$

$$\frac{\partial h}{\partial t}=\frac{\partial\varphi}{\partial\nu}-\varepsilon \quad \text{if } x\in\Gamma, \qquad h(x,t)>0, \tag{0.4}$$

$$\frac{\partial h}{\partial t}=\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^{+} \quad \text{if } x\in\Gamma, \qquad h(x,t)=0, \tag{0.5}$$

$$h(x,0)=0 \quad \text{on } \Gamma, \tag{0.6}$$

where $\varepsilon>0$ is the dissolution current constant; $\partial/\partial\nu$ denotes the inward normal derivative. Let ψ be the solution of (0.1), (0.2), and

$$\psi=0 \quad \text{on } \Gamma. \tag{0.7}$$

Obviously, if $\partial\psi/\partial\nu\leqq\varepsilon$ on Γ, then $(\psi,0)$ is a trivial solution of (0.1)-(0.6). In order to exclude this case, we will always assume in the sequel that

$$\operatorname{meas}\left(\left\{x\in\Gamma:\frac{\partial\psi}{\partial\nu}>\varepsilon\right\}\right)>0. \tag{0.8}$$

* Received by the editors October 1, 1986; accepted for publication (in revised form) December 20, 1987.

† Dipartimento di Matematica, Università di Pavia, Strada Nuova, 65, 27100 Pavia, Italy. The research of this author was done while he was visiting the Laboratoire d'Analyse Numérique, Université Paris VI, Paris, France in 1986 and supported by I.A.N.-C.N.R. and M.P.I., Italy.

‡ Université Pierre et Marie Curie, Laboratoire d'Analyse Numérique, Tour 55-65, 5ème Etage, 4, Place Jussieu, 75230 - Paris Cedex 05, France.

Under this assumption, Caffarelli and Friedman showed in [CF] that a smooth solution of (0.1)-(0.6) satisfies

$$\frac{\partial h}{\partial t}=\left(\frac{\partial \varphi}{\partial \nu}-\varepsilon\right)^{+} \quad \text{on } \Gamma, \tag{0.9}$$

i.e., there is no paint dissolution in the process; moreover, they introduced a time-discretized version of the above problem with (0.9) instead of (0.4), (0.5) and they proved existence and uniqueness of the discretized solution and its convergence to the steady state Signorini problem.

The study of the electropainting problem has been taken up by Marquez and Shillor in [MS], where they introduced an overpotential $\sigma(x)\geqq\sigma_*>0$ for the paint thickness h. Namely they considered the following conditions:

$$\frac{\partial h}{\partial t}=\frac{\partial \varphi}{\partial \nu}-\varepsilon \quad \text{if } x\in\Gamma, \qquad h(x,t)>\sigma(x), \tag{0.10}$$

$$\frac{\partial h}{\partial t}=\left(\frac{\partial \varphi}{\partial \nu}-\varepsilon\right)^{+} \quad \text{if } x\in\Gamma, \qquad h(x,t)=\sigma(x), \tag{0.11}$$

$$h(x,0)=\sigma(x) \quad \text{if } x\in\Gamma, \tag{0.12}$$

instead of (0.4)-(0.6).

Note that this is an approximation that regularizes the original problem because it avoids degeneracy in the Neuman condition (0.3). Thus they are able to show that a smooth solution of (0.1)-(0.3), (0.10)-(0.12) satisfies (0.9) and, moreover, that $\partial h/\partial t\not\equiv 0$ on Γ is impossible. The work of Marquez and Shillor also contains an existence and uniqueness result for the smooth solution of (0.1)-(0.3), (0.9), (0.12). Furthermore an $L^3(\Gamma)$ estimate of $\partial\varphi/\partial\nu$ is proved in [MS, Lemma 6.3] under the following geometrical assumption: Ω is the difference between a convex set with boundary Γ, and a subset of it with boundary S; thus Γ and S are inverted.

The novelty in this paper is the extension of this $L^3(\Gamma)$ estimate to Ω with a general geometry; such a generalization is obtained via a harmonic supersolution technique.

Then, letting $\sigma\to 0$, we are able to obtain a weak solution of the problem (0.1)-(0.3), (0.9), (0.6) as a monotone limit of the solutions $(\varphi_\sigma, h_\sigma)$ given in [MS] for $\sigma(x)=\sigma>0$. Furthermore, showing that in fact $(\varphi_\sigma, h_\sigma)$ is a solution of (0.1)-(0.3), (0.10)-(0.12), we prove that the obtained weak solution satisfies (0.4), (0.5), i.e., it is a solution of the electropainting problem. Finally, following the argument given in [CF], it can be shown that the process converges asymptotically to a unique steady state.

1. An existence result. Let Ω_1 and Ω_2 be connected bounded open sets in $\mathbb{R}^N$ ($N\geqq 2$) with $C^{1,1}$ boundaries Γ and S, respectively. We assume that $\bar{\Omega}_1\subset\Omega_2$ and set $\Omega:=\Omega_2\backslash\bar{\Omega}_1$. In the physical situation Ω_1 is the workpiece and Ω_2 is the bath.

We consider the following problem:

(P_0) Given $T>0$, find a couple $\{\varphi, h\}$, $\varphi:\Omega\times[0,T]\to\mathbb{R}$, $h:\Gamma\times[0,T]\to\mathbb{R}$ satisfying

$$\Delta\varphi=0 \quad \text{in } \Omega, \qquad 0\leqq t\leqq T, \tag{1.1}$$

$$\varphi=1 \quad \text{on } S, \qquad 0\leqq t\leqq T, \tag{1.2}$$

$$h\frac{\partial \varphi}{\partial \nu}=\varphi \quad \text{on } \Gamma, \qquad 0\leqq t\leqq T, \tag{1.3}$$

$$h=\int_0^t\left(\frac{\partial \varphi}{\partial \nu}-\varepsilon\right)^{+} d\tau \quad \text{on } \Gamma, \qquad 0\leqq t\leqq T, \tag{1.4}$$

where $\partial\varphi/\partial\nu$ is the inward normal derivative of φ on Γ and $\varepsilon > 0$ is a given real number.

First we define a weak solution of (P_0).

DEFINITION 1.1. A weak solution of (P_0) is a couple of functions $\{\varphi, h\}$ such that

$$\varphi \in L^\infty(0, T; H^{3/2}(\Omega)), \tag{1.5}$$

$$h \in L^\infty(\Gamma \times (0, T)), \qquad h_t \in L^\infty(0, T; L^2(\Gamma)), \tag{1.6}$$

where h_t is the derivative of h with respect to t;

$$\varphi \text{ and } h \text{ satisfy (1.1)-(1.4).} \tag{1.7}$$

Note that the equalities (1.3), (1.4) make sense; indeed from (1.5) and (1.1) it follows that $\partial\varphi/\partial\nu \in L^\infty(0, T; L^2(\Gamma))$ (cf., e.g., [LM, Chap. 2, Thm. 7.3]).

Our aim is to prove the existence of a weak solution of (P_0). To this end, we recall a result of Marquez and Shillor [MS] on the existence and uniqueness of a smooth solution of the following problem:

(P_σ) Find $(\varphi_\sigma, h_\sigma)$ satisfying (1.1)-(1.3) and

$$h_\sigma = \sigma + \int_0^t \left(\frac{\partial\varphi_\sigma}{\partial\nu} - \varepsilon\right)^+ d\tau \quad \text{on } \Gamma, \qquad 0 \leqq t \leqq T, \tag{1.8}$$

where $\sigma > 0$ is a constant.

THEOREM 1.1. *There exists one and only one solution φ_σ, h_σ of (P_0) such that*

$$\varphi_\sigma \in W^{1,\infty}(0, T; H^{3/2}(\Omega)) \cup C^{0,\alpha}(\bar{\Omega} \times [0, T]), \tag{1.9}$$

$$\nabla\varphi_\sigma \in C^0(\bar{\Omega} \times [0, T]), \tag{1.10}$$

$$(\varphi_\sigma)_t \in C^\alpha(\bar{\Omega} \times [0, T]), \tag{1.11}$$

$$\varphi_\sigma, (\varphi_\sigma)_t \in C^{1,\alpha}(\bar{\Omega}), \qquad 0 \leqq t \leqq T, \tag{1.12}$$

$$\frac{\partial\varphi_\sigma}{\partial\nu}, \left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t \in C^\alpha(\Gamma \times [0, T]), \tag{1.13}$$

$$h_\sigma, (h_\sigma)_t \in C^\alpha(\Gamma \times [0, T]), \tag{1.14}$$

for any $\alpha \in (0, 1)$.

Proof. See Theorems 4.2 and 4.4 of [MS].

We will use these results to prove the next theorem.

THEOREM 1.2. *There exists at least one weak solution of (P_0).*

Before giving the proof, we state some preliminary results. The next lemma generalizes an idea of [MS] for a particular geometry of Ω.

LEMMA 1.1. *Let W be the solution of*

$$\begin{aligned} &\Delta W = 0 && \text{in } \Omega, \\ &W = 1 && \text{on } S, \\ &\sigma\frac{\partial W}{\partial\nu} = W && \text{on } \Gamma. \end{aligned} \tag{1.15}$$

Then there is a constant $\theta > 0$ independent of $\sigma > 0$ such that

$$0 \leqq \frac{\partial W}{\partial\nu} \leqq \theta. \tag{1.16}$$

Proof. As Γ is $C^{1,1}$, Ω_1 has the uniform interior sphere property, i.e., there exists $r_0>0$ such that for every $x\in\Gamma$ there exists a sphere of radius r_0 contained in Ω_1 and tangent to Γ at the point x.

Denote by x_0 the point of Γ where the maximum of W on Γ is achieved. Then the maximum of $\partial W/\partial\nu$ on Γ is also achieved at x_0 (cf. (1.15)). We want to find a supersolution U such that $(\partial U/\partial\nu)(x_0)\geqq(\partial W/\partial\nu)(x_0)$; then, in order to prove the lemma, it will be sufficient to bound $(\partial U/\partial\nu)(x_0)$.

Let y be the center of the ball of radius r_0 contained in Ω_1 and tangent to Γ in x_0. Set

$$r=|x-y|,\qquad d_0=d(S,\Gamma),$$
$$\text{if}\quad N\geqq 3\quad U(x)=C_1-C_2/r^{N-2},$$
$$\text{if}\quad N=2\quad U(x)=C_1-C_2\log r,$$

where C_1 and C_2 are given by

$$\begin{aligned} &C_1-C_2/r_0^{N-2}=W(x_0)\\ &C_1-C_2/(r_0+d_0)^{N-2}=1 \end{aligned}\qquad \text{if}\quad N\geqq 3, \tag{1.17}$$

and, respectively,

$$\begin{aligned} &C_1-C_2\log r_0=W(x_0)\\ &C_1-C_2\log(r_0+d_0)=1 \end{aligned}\qquad \text{if}\quad N=2. \tag{1.18}$$

Note that U satisfies

$$\begin{aligned} &\Delta U=0 && \text{in } \Omega,\\ &U\geqq 1 && \text{on } S,\\ &U\geqq W && \text{on } \Gamma. \end{aligned} \tag{1.19}$$

By the comparison principle, from (1.15) and (1.19) it follows that $U\geqq W$ on $\bar\Omega$. Moreover $U(x_0)=W(x_0)$; therefore by the strong maximum principle we infer that $(\partial U/\partial\nu)(x_0)\geqq(\partial W/\partial\nu)(x_0)$.

It is easy to check that C_2 is bounded independently of σ and x_0 (note that $0<W(x)<1$ for $x\in\Omega\cup\Gamma$) and this gives the desired estimate for $(\partial U/\partial\nu)(x_0)$.

LEMMA 1.2. *Let $(\varphi_\sigma, h_\sigma)$ be the solution of (P_σ) given by Theorem* 1.1. *Then there exists a constant $C>0$ independent of σ, T such that*

$$\int_\Gamma\left|\frac{\partial\varphi_\sigma}{\partial\nu}\right|^3\leqq C \quad \textit{for any } \sigma>0,\ t\in[0,T]. \tag{1.20}$$

Proof. See Lemma 6.3 of [MS]. However, since their lemma is the crucial step needed in passing to the limit with $\sigma\to 0$, we briefly outline their proof.

By assumption, $(\varphi_\sigma, h_\sigma)$ satisfies

$$h_\sigma\frac{\partial\varphi_\sigma}{\partial\nu}=\varphi_\sigma \quad \text{on } \Gamma,\ 0\leqq t\leqq T. \tag{1.3}$$

Differentiating (1.3) with respect to time, then multiplying it by $(\partial\varphi_\sigma/\partial\nu)_t$, integrating over Γ and noting that

$$\int_\Gamma(\varphi_\sigma)_t\left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t=-\int_\Omega|\nabla(\varphi_\sigma)_t|^2\leqq 0,$$

and that if $\partial\varphi_\sigma/\partial\nu \leqq \varepsilon$ then $(\partial\varphi_\sigma/\partial\nu)_t > 0$, we obtain

$$\int_\Gamma \left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)\frac{\partial\varphi_\sigma}{\partial\nu}\left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t \leqq \int_\Gamma \left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+\frac{\partial\varphi_\sigma}{\partial\nu}\left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t$$

$$\leqq \int_\Gamma (\varphi_\sigma)_t\left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t \leqq 0,$$

which can be rewritten as

$$\int_\Gamma \left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^2\left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t + \varepsilon\int_\Gamma \left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)\left(\frac{\partial\varphi_\sigma}{\partial\nu}\right)_t \leqq 0;$$

i.e.,

$$\frac{d}{dt}\int_\Gamma \left\{\frac{1}{3}\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^3+\frac{\varepsilon}{2}\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^2\right\} \leqq 0.$$

The estimate follows by integration over $(0, t)$ and from the uniform estimate of $(\partial\varphi_\sigma/\partial\nu)\ (x, 0)$ given by Lemma 1.1.

Now we collect some results proved in [MS], namely Lemmas 6.4, 4.3, and 6.5.

LEMMA 1.3. *There exists $C > 0$ independent of σ and T such that*

$$\|\varphi_\sigma\|_{H^1(\Omega)} \leqq C, \qquad 0 \leqq t \leqq T. \tag{1.21}$$

Moreover for every $\sigma < 1/\varepsilon$

$$0 \leqq h_\sigma \leqq 1+\frac{1}{\varepsilon} \quad \text{on } \Gamma\times[0, T]. \tag{1.22}$$

Finally if $0 < \sigma_1 < \sigma_2$ and $(\varphi_{\sigma_1}, h_{\sigma_1})$, $(\varphi_{\sigma_2}, h_{\sigma_2})$ are, respectively, the solutions of (P_{σ_1}) and (P_{σ_2}) given by Theorem 1.1, then

$$\varphi_{\sigma_1} \leqq \varphi_{\sigma_2} \quad \text{on } (\Omega\cup\Gamma)\times[0, T], \tag{1.23}$$

$$h_{\sigma_1} \leqq h_{\sigma_2} \quad \text{on } \Gamma\times[0, T]. \tag{1.24}$$

Proof of Theorem 1.2. Let us denote by φ, h the monotone limits of the sequences φ_σ, h_σ as $\sigma \to 0$ (cf. (1.23), (1.24)). From (1.21), (1.22) it follows that

$$\varphi_\sigma \searrow \varphi \quad \text{weakly in } H^1(\Omega), \quad 0 < t < T, \quad \text{a.e. in } \Omega, \tag{1.25}$$

$$h_\sigma \searrow h \quad \text{weakly * in } L^\infty(\Gamma\times(0, T)), \quad \text{a.e. in } \Gamma. \tag{1.26}$$

From (1.20), (1.21), and (1.1) for φ_σ it follows that $\{\varphi_\sigma\}$ actually is bounded in $L^\infty(0, T; H^{3/2}(\Omega))$ (see, e.g., [LM, Chap. 2, (7.28)]); thus $\varphi_\sigma \to \varphi$ weakly * in this space, which yields $\partial\varphi_\sigma/\partial\nu \to \partial\varphi/\partial\nu$ weakly * in $L^\infty(0, T; L^2(\Gamma))$ (see for instance [LM, Chap. 2, Thm. 7.3]); by (1.20) normal derivatives actually converge in $L^\infty(0, T; L^3(\Gamma))$. Note that all this is true in principle for a subsequence, but the monotonicity of $\{\varphi_\sigma\}$ implies a unique limit and thus the "whole" sequences converge.

Due to Lebesgue's theorem, we have that $h_\sigma \to h$ strongly in $L^p(\Gamma)$ for any $1 < p < +\infty$, $0 \leqq t \leqq T$. Then, taking the limit in

$$h_\sigma\frac{\partial\varphi_\sigma}{\partial\nu} = \varphi_\sigma,$$

we obtain (1.3). Note that h is a nondecreasing function with respect to t, as it is the limit of nondecreasing functions. It remains to prove (1.4); therefore it suffices to show that

$$\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+ \xrightarrow[\sigma\to 0]{} \left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+ \quad \text{in } L^1(\Gamma\times(0,T)). \tag{1.27}$$

Indeed (1.27) will allow us to take the limit in (1.8) to obtain (1.4).

Set the following:

$$\begin{aligned} &Q_0=\{(x,t)\in\Gamma\times(0,T): h(x,t)=0\},\\ &Q_+=(\Gamma\times(0,T))\setminus Q_0, \end{aligned} \tag{1.28}$$

$$\begin{aligned} &\Gamma_0(t)=\{x\in\Gamma: h(x,t)=0\},\\ &\Gamma_+(t)=\Gamma\setminus\Gamma_0(t). \end{aligned} \tag{1.29}$$

As h is monotone nondecreasing with respect to t, $\Gamma_0(t)\times[0,t]\subset Q_0$ for any $t\in[0,T]$. Now, integrating (1.8) on $\Gamma_0(t)$ and taking the limit as $\sigma\to 0$, we get

$$\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+\to 0 \quad \text{in } L^1(\Gamma_0(t)\times[0,t]), \qquad 0\leqq t\leqq T. \tag{1.30}$$

As $\partial\varphi_\sigma/\partial\nu\to\partial\varphi/\partial\nu$ weakly * in $L^\infty(0,T;L^3(\Gamma))$, thus weakly in $L^1(\Gamma_0(t)\times[0,T])$ for any $t\in[0,T]$, from the equality

$$\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)=\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+-\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^- \tag{1.31}$$

and (1.30) it follows that the weak limit $\xi:=\lim((\partial\varphi_\sigma/\partial\nu)-\varepsilon)^-$ exists in $L^1(\Gamma_0(t)\times[0,T])$ and that $(\partial\varphi/\partial\nu)-\varepsilon=0-\xi\leqq 0$; that is,

$$\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+=0 \quad \text{on } \Gamma_0(t)\times[0,t],\ 0\leqq t\leqq T. \tag{1.32}$$

Now, because of the monotonicity of the set $\Gamma_0(t)$, it is possible to find a sequence $t_m\in(0,T)$ such that

$$\text{meas}\,(Q_0-E_n)\to 0 \quad \text{as } n\to\infty, \tag{1.33}$$

where $E_n=U_{m\leqq n}\Gamma_0(t_m)\times(0,t_m)$. Note that (1.30) and (1.32) are still true in $L^1(E_n)$.

Applying the Schwarz-Hölder inequality and (1.20), we have

$$\begin{aligned} \int_{Q_0}\left|\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+-\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+\right| &= \int_{Q_0\setminus E_n}\left|\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+-\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+\right|+\int_{E_n}\left|\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+\right|\\ &\leqq C\{\text{meas}\,(Q_0-E_n)\}^{2/3}+\int_{E_n}\left|\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+\right|, \end{aligned} \tag{1.34}$$

where C is independent of σ. Taking the lim sup as $\sigma\to 0$ in (1.34), we obtain by (1.30)

$$\limsup_{\sigma\to 0}\int_{Q_0}\left|\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+-\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+\right|\leqq C\{\text{meas}\,(Q_0-E_n)\}^{2/3}. \tag{1.35}$$

From (1.33), (1.35) it follows that (1.27) is satisfied on Q_0.

In addition, by the definition of Q_+, (1.25), (1.26), and (1.3), we have

$$\frac{\partial \varphi_\sigma}{\partial \nu} = \frac{\varphi_\sigma}{h_\sigma} \xrightarrow[\sigma \to 0]{} \frac{\varphi}{h} = \frac{\partial \varphi}{\partial \nu} \quad \text{a.e. on } Q_+. \tag{1.36}$$

Now Egorov's theorem asserts that for every $\delta > 0$ there exists a subset $A_\delta \subset Q_+$ such that $\partial \varphi_\sigma / \partial \nu \longrightarrow_{\sigma \to 0} \partial \varphi / \partial \nu$ uniformly on A_δ and meas $(Q_+ \setminus A_\delta) < \delta$. Then, using the Schwarz-Hölder inequality, we derive from (1.20)

$$\int_{Q_+} \left| \frac{\partial \varphi_\sigma}{\partial \nu} - \frac{\partial \varphi}{\partial \nu} \right| \leqq C\delta^{2/3} + \int_{A_\delta} \left| \frac{\partial \varphi_\sigma}{\partial \nu} - \frac{\partial \varphi}{\partial \nu} \right|, \tag{1.37}$$

where C is independent of σ. Thus (1.27) is also satisfied on Q_+. This concludes the proof of Theorem 1.2. □

Remark 1.1. Note that, by Theorem 1.2, problem (P_0) has a solution for every $T > 0$, i.e., on $[0, +\infty)$.

2. A solution of the electropainting problem. In [MS], Marquez and Shillor, following Caffarelli and Friedman [CF], show a smooth solution (in the sense of Theorem 1.1) of the following problem:

(EP_σ) Find $(\varphi_\sigma, h_\sigma)$ satisfying (1.1)-(1.3) and

$$(h_\sigma)_t = \left(\frac{\partial \varphi_\sigma}{\partial \nu} - \varepsilon \right)^+ \quad \text{on } \Gamma \cap \{h_\sigma = \sigma\}, \qquad t \geqq 0, \tag{2.1}$$

$$(h_\sigma)_t = \left(\frac{\partial \varphi_\sigma}{\partial \nu} - \varepsilon \right) \quad \text{on } \Gamma \cap \{h_\sigma > \sigma\}, \qquad t \geqq 0, \tag{2.2}$$

$$h_\sigma(x, 0) = \sigma \quad \text{on } \Gamma, \tag{2.3}$$

where $\sigma > 0$ is a smooth solution of (P_σ).

Our aim is to prove that the converse is also true, by means of an argument suggested by Theorem 2.1 of [CF]. This result will allow us to take the limit in (2.1)-(2.3) as $\sigma \to 0$ in order to obtain a solution of the following problem:

(EP_0) Find (φ, h) satisfying (1.1)-(1.3) and

$$h_t = \left(\frac{\partial \varphi}{\partial \nu} - \varepsilon \right)^+ \quad \text{on } \Gamma \cap \{h = 0\}, \qquad t \geqq 0, \tag{2.4}$$

$$h_t = \left(\frac{\partial \varphi}{\partial \nu} - \varepsilon \right) \quad \text{on } \Gamma \cap \{h > 0\}, \qquad t \geqq 0, \tag{2.5}$$

$$h(x, 0) = 0 \quad \text{on } \Gamma. \tag{2.6}$$

THEOREM 2.1. *There exists a unique solution* $(\varphi_\sigma, h_\sigma)$ *of* (EP_σ) *satisfying* (1.9)-(1.14).

Proof. Let $(\varphi_\sigma, h_\sigma)$ be the solution of (P_σ) given by Theorem 1.1. In order to prove that $(\varphi_\sigma, h_\sigma)$ is a solution of (EP_σ), it suffices to show that h_σ verifies (2.2). By contradiction, we assume that there exists $x_0 \in \Gamma$ and $c > \sigma$ such that

$$h_\sigma(x_0, t) = c \tag{2.7}$$

for any t belonging to an open interval $I \subset (0, +\infty)$. Let $(t_0, \bar{t})$ be the maximal interval where (2.7) is satisfied. Note that, by (2.3), $t_0 > 0$.

Step 1. $(\varphi_\sigma)_t \geqq 0$. Since h_σ is monotone increasing, the claim follows from the following property (see [CF, Lemma 2.2]): if φ^1, φ^2 are the solutions of

$$\begin{aligned} &\Delta\varphi^i = 0 && \text{in } \Omega, \\ &\varphi^i = 1 && \text{on } S, \\ &h_i \frac{\partial \varphi_i}{\partial \nu} = \varphi_i && \text{on } \Gamma, \quad i = 1, 2, \end{aligned} \tag{2.8}$$

corresponding, respectively, to h_1, h_2, and if $h_1 \geqq h_2$ on Γ, then $\varphi_1 \geqq \varphi_2$ in Ω.

Step 2. $(\partial\varphi_\sigma/\partial\nu)_t(x_0, t_0) \leqq 0$, $(h_\sigma)_t(x_0, t_0) = 0$. Since $(t_0, \bar{t})$ is the maximal interval where (2.7) is satisfied, by (1.8) there exists $t_i \nearrow t_0$ such that $(\partial\varphi_\sigma/\partial\nu)(x_0, t_i) > \varepsilon$. On the other hand, $\partial\varphi_\sigma/\partial\nu(x_0, t_0) \leqq \varepsilon$. Then $(\partial\varphi_\sigma/\partial\nu)_t(x_0, t_0) \leqq 0$. $(h_\sigma)_t(x_0, t_0) = 0$ follows from (2.7).

Step 3. $(\varphi_\sigma)_t(x, t_0) = 0$ for $x \in \Omega$, $(h_\sigma)_t(x, t_0) = 0$ for $x \in \Gamma$. The expression $(\varphi_\sigma)_t$ is a solution of the following:

$$\Delta(\varphi_\sigma)_t = 0 \quad \text{in } \Omega, \tag{2.9}$$

$$(\varphi_\sigma)_t = 0 \quad \text{on } S, \tag{2.10}$$

$$(h_\sigma)_t \frac{\partial \varphi_\sigma}{\partial \nu} + h_\sigma \frac{\partial}{\partial \nu}(\varphi_\sigma)_t = (\varphi_\sigma)_t \quad \text{on } \Gamma. \tag{2.11}$$

From (2.7), Step 1, Step 2, and (2.11) we find

$$(\varphi_\sigma)_t(x_0, t_0) = \frac{\partial}{\partial \nu}(\varphi_\sigma)_t(x_0, t_0) = 0. \tag{2.12}$$

Then, by the strong maximum principle, $(\varphi_\sigma)_t(x, t_0) = 0$ for every $x \in \Omega$; therefore, $(\partial/\partial\nu)(\varphi_\sigma)_t(x, t_0) = 0$ for $x \in \Gamma$ and (2.11) becomes

$$(h_\sigma)_t \frac{\partial \varphi_\sigma}{\partial \nu} = 0 \quad \text{on } \Gamma.$$

It follows that $(h_\sigma)_t(x, t_0) = 0$ for $x \in \Gamma$ ($(\partial\varphi_\sigma/\partial\nu) > 0$, by the strong maximum principle applied to (1.1)-(1.3)).

Step 4. Definition of τ. For any $x \in \Gamma$ set

$$A(x) = \bigcup_{\alpha \in \mathcal{A}} I_\alpha(x), \tag{2.13}$$

where $I_\alpha(x)$ is an open interval of $(0, +\infty)$ such that

$$h(x, t) = C_\alpha > \sigma \quad \text{for every} \quad t \in I_\alpha, \tag{2.14}$$

where C_α is a constant and $\mathcal{A}$ indexes all the intervals satisfying (2.14). Set

$$t_0(x) = \begin{cases} \inf A(x) & \text{if } A(x) \neq \varnothing, \\ +\infty & \text{if } A(x) = \varnothing. \end{cases} \tag{2.15}$$

Now, let τ be defined by

$$\tau = \inf_{x \in \Gamma} t_0(x). \tag{2.16}$$

Note that $\tau < +\infty$ by our contradiction assumption.

Step 3 implies $(h_\sigma)_t(x, t_0(y)) = 0$ for every $x, y \in \Gamma$ with $t_0(y) < +\infty$, and consequently $(h_\sigma)_t(x, \tau) = 0$ for $x \in \Gamma$. Hence $\tau > 0$, if not, we would have a contradiction with (0.8).

Step 5. End of the proof. By the definition of τ, we have that for $t<\tau$ the solution of (P_σ) is a solution of (EP_σ). Therefore we can apply the final argument of the proof of Theorem 2.3 in [MS] to obtain a contradiction.

The uniqueness of the solution is obvious for the equivalence between (P_σ) and (EP_σ) and the uniqueness of the solution of (P_σ).

THEOREM 2.2. *There exists a weak solution (in the sense of Definition* 1.1*) of problem* (EP_0).

Proof. Let (φ, h) be the solution of (P_0) obtained in § 1 as the limit of $(\varphi_\sigma, h_\sigma)$. We want to show that in fact (φ, h) is a solution of (EP_0). The only thing to prove is (2.5).

We fix $T>0$. Due to the monotonicity of the set $\Gamma_+(t)$, it is possible to find a sequence $t_m \in (0, T)$ such that, if $A_n = \bigcup_{m\leqq n} \Gamma_+(t_m)\times(t_m, T)$, then

$$\text{meas}\,(Q_+ \setminus A_n)\to 0 \quad \text{as} \quad n\to+\infty. \tag{2.17}$$

Set

$$B_n = \bigcup_{\substack{m\leqq n\\ l\leqq n}} \left\{(x,t): h(x,t)>\frac{1}{m},\ t_l\leqq t\leqq T\right\}. \tag{2.18}$$

Note that A_n and B_n are two monotone set sequences and

$$\bigcup_{n\in\mathbb{N}} B_n = \bigcup_{n\in\mathbb{N}} A_n \tag{2.19}$$

so that

$$\text{meas}\left(Q_+ \setminus \left(\bigcup_{n\in\mathbb{N}} B_n\right)\right)=0. \tag{2.20}$$

Using (1.24) and (1.26) we have

$$h_\sigma(x,t)\geqq h(x,t)>\frac{1}{n}>\sigma \quad \text{for } (x,t)\in B_n,\ \ \sigma<\frac{1}{n}. \tag{2.21}$$

From Theorem 2.1 it follows that

$$(h_\sigma)_t=\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon=\left(\frac{\partial\varphi_\sigma}{\partial\nu}-\varepsilon\right)^+ \quad \text{on } B_n. \tag{2.22}$$

As $\partial\varphi_\sigma/\partial\nu\to\partial\varphi/\partial\nu$ weakly in $L^3(\Gamma)$ for every $t\in[0,T]$ and $((\partial\varphi_\sigma/\partial\nu)-\varepsilon)^+\to((\partial\varphi/\partial\nu)-\varepsilon)^+$ in $L^1(\Gamma\times[0,T])$ as $\sigma\to 0$ (cf. (1.20) and (1.27)), we can deduce from (2.22)

$$\frac{\partial\varphi}{\partial\nu}-\varepsilon=\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+ \quad \text{on } B_n. \tag{2.23}$$

Finally from (2.20) it follows that

$$\frac{\partial\varphi}{\partial\nu}-\varepsilon=\left(\frac{\partial\varphi}{\partial\nu}-\varepsilon\right)^+ \quad \text{on } Q_+. \tag{2.24}$$

Since T is arbitrary, (2.5) is proved.

Remark 2.1. Concerning the asymptotic behaviour of the solution (φ, h) of the electropainting problem (EP_0), we recall that φ and h increase in t and are bounded in $H^1(\Omega)$ and $L^\infty(\Gamma)$, respectively. We can thus define

$$w:=\lim_{t\to\infty}\varphi(\cdot,t) \quad \text{weakly in } H^1(\Omega), \tag{2.25}$$

$$\bar{h}:=\lim_{t\to\infty} h(\cdot,t) \quad \text{weakly * in } L^\infty(\Gamma). \tag{2.26}$$

We are able to prove that w is the unique solution to the following variational inequality:

$$
\begin{aligned}
&w \in K, \quad \int_{\Omega} \nabla w \nabla(\eta - w) + \varepsilon \int_{\Gamma} (\eta - w) \geqq 0 \quad \forall \eta \in K, \\
&K = \{\eta \in H^1(\Omega): \quad \eta = 1 \quad \text{on } S, \quad \eta \geqq 0 \quad \text{on } \Gamma\}.
\end{aligned} \tag{2.27}
$$

For the proof, it is sufficient to follow the arguments of [CF, § 5]. Then the electropainting process stabilizes as $t \to \infty$ and tends asymptotically to a unique steady state; in fact w is a $C^{1,\alpha}$-function (cf. [F]) and

$$
\bar{h}(x) = \begin{cases} w(x) \Big/ \dfrac{\partial w}{\partial \nu}(x) & \text{if} \quad w(x) > 0, \\ 0 & \text{if} \quad w(x) = 0. \end{cases} \tag{2.28}
$$

REFERENCES

[ALS] J. M. AITCHISON, A. A. LACEY, AND M. SHILLOR, *A model for an electropaint process*, IMA J. Appl. Math., 33 (1984), pp. 17-31.

[CF] L. A. CAFFARELLI AND A. FRIEDMAN, *A nonlinear evolution problem associated with an electropaint process*, SIAM J. Math. Anal., 16 (1985), pp. 955-969.

[F] A. FRIEDMAN, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[LM] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, Berlin, 1972.

[MS] V. MARQUEZ AND M. SHILLOR, *The electropainting problem with overpotentials*, SIAM J. Math. Anal., 18 (1987), pp. 788-811.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

005

EXACT ESTIMATES FOR POTENTIALS*

MARTIN SCHECHTER†

Abstract. For $\lambda \geqq 0$ it is determined precisely which functions $V(x) \geqq 0$ on $\mathbf{R}^n$ satisfy an inequality of the form

$$(Vu, u) \leqq C_\lambda(V)(\|\nabla u\|^2 + \lambda^2 \|u\|^2), \qquad u \in C_0^\infty$$

for some constant $C_\lambda(V)$. The value of the smallest such constant is found. Inequalities of this type are important in the study of the Schrödinger equation. An application is given.

Key words. Bessel potentials, Schrödinger operators, spectral theory

AMS(MOS) subject classifications. primary 26D10, 45P05, 47G05; secondary 26O15, 35P15, 35J10

1. Introduction.

Let

$$(u, v) = \int_{\mathbf{R}^n} u(x)v(x)^* \, dx, \qquad \|u\| = (u, u)^{1/2}$$

denote the scalar product and norm in $L^2 = L^2(\mathbf{R}^n)$, and let $V(x) \geqq 0$ be a measurable function on $\mathbf{R}^n$. For $\lambda \geqq 0$ define

$$C_\lambda(V) = \sup_{u \in C_0^\infty} \frac{(Vu, u)}{\|\nabla u\|^2 + \lambda^2 \|u\|^2} \tag{1}$$

where C_0^∞ denotes the set of all infinitely differentiable functions on $\mathbf{R}^n$ with compact supports. Our first result is Theorem 1.

THEOREM 1.

$$C_\lambda(V) = \inf_{\rho > 0} \sup_y \frac{1}{\rho(y)} \int V(x)\rho(x)G_\lambda(x-y) \, dx \tag{2}$$

where

$$G_\lambda(x) = \frac{|x|^{2-n}}{4\pi^{n/2}} \int_0^\infty e^{-(\lambda^2|x|^2/4t)-t} t^{n/2-2} \, dt \tag{3}$$

(*if* $n \leqq 2$, *assume* $\lambda \neq 0$).

As a consequence we have Corollary 2.

COROLLARY 2. *If* $n > 2$, *then*

$$C_0(V) = \frac{\Gamma(n/2-1)}{4\pi^{n/2}} \inf_{\rho > 0} \sup_y \frac{1}{\rho(y)} \int V(x)\rho(x)|x-y|^{2-n} \, dx. \tag{4}$$

Thus there is a constant C *such that*

$$(Vu, u) \leqq C\|\nabla u\|^2, \qquad u \in C_0^\infty \tag{5}$$

if and only if for each $\varepsilon > 0$ *there is a* $\rho(x) > 0$ *such that*

$$\int V(x)\rho(x)|x-y|^{2-n} \, dx \leqq \frac{4\pi^{n/2}}{\Gamma(\frac{1}{2}n-1)} (C+\varepsilon)\rho(y). \tag{6}$$

For other work on this question see [2], [3], [5]–[10] and the references therein.

* Received by the editors February 10, 1986; accepted for publication February 24, 1988.
† Department of Mathematics, University of California, Irvine, CA 92717.

2. Integral operators. Our proof of Theorem 1 will be based on Theorem 3.

THEOREM 3. *Let $K(x, y) \geqq 0$ be a measurable function on $\mathbf{R}^{2n}$, and define*

$$Tu(x) = \int K(x, y)u(y)\, dy. \tag{7}$$

Then T is a bounded operator on $L^2(\mathbf{R}^n)$ if and only if there are a function $\phi(x) \geqq 0$ on $\mathbf{R}^n$ and a constant C such that

(a) $K(x, y) = 0$ *a.e. when* $\psi(x) = 0$ *and* $\phi(y) = 0$

and

(b) $\int K(x, y)\psi(x)\, dx \leqq C\phi(y)$ *a.e.*

where

$$\psi(x) = \int K(x, y)\phi(y)\, dy. \tag{8}$$

Moreover, $\|T\|^2 \leqq C$. If

$$C_0 = \inf_{\phi} \sup_{y} \frac{1}{\phi(y)} \int K(x, y)K(x, z)\phi(z)\, dz\, dx \tag{9}$$

where the infimum is taken over all $\phi(x) \geqq 0$ satisfying (a) *and* (b) *and we take $0/0 = 0$, then*

$$\|T\|^2 = C_0. \tag{10}$$

Proof. Assume that there is a $\phi(z) \geqq 0$ *satisfying* (a) and (b). Let

$$M = \{x \in R^n \mid \psi(x) = 0\}, \qquad N = \{x \in R^n \mid \phi(x) = 0\},$$
$$M' = R^n - M, \qquad N' = R^n - N.$$

Thus we have

$$\begin{aligned} K(x, y) &= 0 \quad \text{a.e.} \quad x \in M, \quad y \in N \quad \text{by (a)}, \\ &= 0 \quad \text{a.e.} \quad x \in M', \quad y \in N \quad \text{by (b)}, \\ &= 0 \quad \text{a.e.} \quad x \in M, \quad y \in N' \quad \text{by (8)}. \end{aligned}$$

Hence

$$(Tu, v) = \int \int K(x, y)u(y)v(x)^*\, dx\, dy$$
$$= \int_{M'} \int_{N'} K(x, y)^{1/2}\psi(x)^{1/2}\phi(y)^{-1/2}u(y)K(x, y)^{1/2}\phi(y)^{1/2}\psi(x)^{-1/2}v(x)^*\, dx\, dy$$

and

$$|(Tu, v)|^2 \leqq \int_{M'} \int_{N'} K(x, y)\psi(x)\phi(y)^{-1}|u(y)|^2\, dx\, dy$$
$$\cdot \int_{M'} \int_{N'} K(x, y)\phi(y)\psi(x)^{-1}|v(x)|^2\, dx\, dy$$
$$\leqq C\|u\|^2\|v\|^2.$$

Thus T is a bounded operator on L^2 and $\|T\|^2 \leqq C$. Since C_0 is the infimum of all constants C satisfying (b), we see that $\|T\|^2 \leqq C_0$.

Conversely, assume that T is bounded on L^2, and let C be any constant satisfying

$$\|T^*T\| < C. \tag{11}$$

Let $h(x)$ be any positive function in L^2, and define

$$\phi_0 = 0, \qquad \phi_{k+1} = h + C^{-1}T^*T\phi_k.$$

Clearly $\phi_k(x) > 0$ for each k and $\phi_k \to \phi$ in L^2. Then

$$\phi = h + C^{-1}T^*T\phi,$$

and consequently $\phi(x) \geqq h(x) > 0$ almost everywhere. Clearly ϕ satisfies (a) and (b). Consequently, any constant C satisfying (11) satisfies (b). Since C_0 is the infimum of such constants, we have $C_0 \leqq \|T^*T\| \leqq \|T\|^2$. □

The proof of Theorem 3 implies

$$\|T\|^2 = \inf_{\phi>0} \sup_y \frac{1}{\phi(y)} \int \int K(x, y)K(x, z)\phi(z)\, dz\, dx. \tag{12}$$

3. The reduction. Now we show how the proof of Theorem 1 can be based on Theorem 3. We shall need the following facts concerning Bessel potentials (cf., e.g., [1], [3], [4]). Let u be any function in C_0^∞, and let

$$v = (\lambda^2 - \Delta)^{1/2}u \tag{13}$$

(v can be defined by Fourier transforms). Then

$$\|v\|^2 = ([\lambda^2 - \Delta]u, u) = \|\nabla u\|^2 + \lambda^2\|u\|^2. \tag{14}$$

Moreover, we have

$$u(x) = \int G_{1,\lambda}(x - y)v(y)\, dy \tag{15}$$

where

$$G_{1,\lambda}(x) = \frac{|x|^{1-n}}{2\pi^{n/2}\Gamma(\frac{1}{2})} \int_0^\infty e^{-(\lambda^2|x|^2/4t)-t} t^{(n-1)/2-1}\, dt. \tag{16}$$

We also have

$$G_{1,\lambda}^* G_{1,\lambda} = G_\lambda \tag{17}$$

where G_λ is given by (3). Using these facts we can give the proof of Theorem 1.

Proof of Theorem 1. Let $\varepsilon > 0$ be given, and let $C_\lambda^*(V)$ denote the right-hand side of (2). If $C_\lambda^*(V) < \infty$, there is a $\rho > 0$ such that

$$\int V(z)\rho(z)G_\lambda(y - z)\, dz \leqq (C_\lambda^*(V) + \varepsilon)\rho(y).$$

Put $\phi(z) = V(z)^{1/2}\rho(z)$, $K(x, y) = G_{1,\lambda}(x - z)V(y)^{1/2}$. Then

$$\psi(x) = \int G_{1,\lambda}(x - z)V(z)\rho(z)\, dz$$

and by (17)

$$\int K(x,y)\psi(x)\,dx = \int G_{1,\lambda}(x-y)\,V(y)^{1/2}\int G_{1,\lambda}(x-z)\,V(z)\rho(z)\,dz\,dx$$

$$= V(y)^{1/2}\int\left[\int G_{1,\lambda}(x-y)G_{1,\lambda}(x-z)\,dx\right]V(z)\rho(z)\,dz$$

$$= V(y)^{1/2}\int G_{\lambda}(y-z)\,V(z)\rho(z)\,dz$$

$$\leqq (C_{\lambda}^{*}(V)+\varepsilon)\,V(y)^{1/2}\rho(y) = (C_{\lambda}^{*}(V)+\varepsilon)\phi(y).$$

Thus condition (b) of Theorem 3 holds. Moreover, the only way that $\psi(x)$ can vanish for some x is if $V(z)\rho(z)\equiv 0$. Since $\rho(z)\neq 0$ almost everywhere, this implies $V(z)\equiv 0$ almost everywhere. This in turn implies $K(x,y)\equiv 0$ almost everywhere. Hence conditions (a) of Theorem 3 are also satisfied. Thus the operator T given by (7) is bounded on L^2 with $\|T\|^2\leqq C_{\lambda}^{*}(V)+\varepsilon$. Since ε was arbitrary we have $\|T\|^2\leqq C_{\lambda}^{*}(V)$. Thus, by (14) and (15)

$$(Vu,u) = \|T^{*}v\|^2 \leqq \|T^{*}\|^2\|v\|^2$$

$$= \|T\|^2(\|\nabla u\|^2+\lambda^2\|u\|^2)$$

$$\leqq C_{\lambda}^{*}(V)(\|\nabla u\|^2+\lambda^2\|u\|^2).$$

Consequently, $C_{\lambda}(V)\leqq C_{\lambda}^{*}(V)$.

Conversely, assume that

$$(Vu,u)\leqq C(\|\nabla u\|^2+\lambda^2\|u\|^2), \qquad u\in C_0^{\infty}. \tag{18}$$

Then by (14) and (15)

$$\|T^{*}v\|^2 = (Vu,u)\leqq C\|v\|^2.$$

Since the range of $(\lambda^2-\Delta)^{1/2}$ on C_0^{∞} is dense in L^2, we see that T^{*} is a bounded operator on L^2 with norm $\leqq C^{1/2}$. Hence $\|T\|^2\leqq C$. By (12) for each $\varepsilon>0$ there is a $\phi>0$ such that

$$\int\int K(x,y)K(x,z)\phi(z)\,dz\,dx\leqq(\|T\|^2+\varepsilon)\phi(y). \tag{19}$$

This is equivalent to

$$\int G_{\lambda}(y-z)\,V(y)^{1/2}V(z)^{1/2}\phi(z)\,dz\leqq(\|T\|^2+\varepsilon)\phi(y). \tag{20}$$

Define ρ by

$$\rho(y) = \frac{\phi(y)}{V(y)^{1/2}}, \qquad V(y)\neq 0$$

$$= (\|T\|^2+\varepsilon)^{-1}\int G_{\lambda}(y-z)\,V(z)^{1/2}\phi(z)\,dz, \qquad V(y)=0.$$

If $V(y)\neq 0$, we have by (20)

$$\int G_{\lambda}(y-z)\,V(z)\rho(z)\,dz\leqq(\|T\|^2+\varepsilon)\rho(y). \tag{21}$$

If $V(y)=0$, we have by the definition of ρ

$$\int G_\lambda(y-z)V(z)\rho(z)\,dz = \int G_\lambda(y-z)V(z)^{1/2}\phi(z)\,dz$$
$$= (\|T\|^2+\varepsilon)\rho(y).$$

Thus (21) holds for all y. Since $\rho(y)>0$ for all y, we see that

$$C_\lambda^*(V) \leqq \|T\|^2+\varepsilon \leqq C+\varepsilon$$

where C is any constant satisfying (18). Thus $C_\lambda^*(V)\leqq C_\lambda(V)$, and the proof is complete. □

4. An application. We can apply Theorem 1 to the study of the spectral theory of the Schrödinger operator. For instance, we have Theorem 4.

THEOREM 4. *Let $V(x)\geqq 0$ be a measurable function on $\mathbf{R}^n$. If $C_\lambda(V)\leqq 1$, then $-\Delta-V$ has a self-adjoint realization H on $L^2(\mathbf{R}^n)$ having no spectrum below $-\lambda^2$. If $C_\lambda(V)>1$, then every self-adjoint realization has spectrum below $-\lambda^2$.*

Proof. If $C_\lambda(V)\leqq 1$, then $(Vu,u)\leqq([\lambda^2-\Delta]u,u)$, $u\in C_0^\infty$. It is well known that this implies that a self-adjoint realization H of $-\Delta-V$ exists such that $-\lambda^2\|u\|^2\leqq(Hu,u)$ (cf., e.g., [3]). This implies that the spectrum of H is contained in the interval $[-\lambda^2,\infty)$. If $C_\lambda(V)>1$, then there is a $u\in C_0^\infty$ such that $(Vu,u)>([\lambda^2-\Delta]u,u)$. Thus $(Hu,u)<-\lambda^2\|u\|^2$ for any self-adjoint realization of $-\Delta-V$. This means that H has spectrum below $-\lambda^2$. □

REFERENCES

[1] N. ARONSZAJN AND K. T. SMITH, *Theory of Bessell potentials*, Ann. Inst. Fourier (Grenoble), 11 (1961), pp. 385-475.

[2] E. BALSLEV, *The essential spectrum of elliptic differential operators in $L^P(R_n)$*, Trans. Amer. Math. Soc., 116 (1965), pp. 193-217.

[3] M. SCHECHTER, *Spectra of Partial Differential Operators*, North-Holland, Amsterdam, 1986.

[4] A. P. CALDERON, *Lebesgue spaces of differentiable functions and distributions*, in Partial Differential Equations, Proc. Symposia in Pure Mathematics, Vol. 4, American Mathematical Society, RI, 1961, pp. 33-49.

[5] F. STUMMEL, *Singulare elliptische Differential operatoren in Hilbertschen Raumen*, Math. Ann., 13 (1956), pp. 150-176.

[6] S. Y. A. CHANG, J. WILSON, AND T. WOLFF, *Some weighted norm inequalities concerning the Schrödinger operators*, Comment Math. Helv., 60 (1985), pp. 217-246.

[7] S. CHANILLO AND R. L. WHEEDEN, *L^P estimates for fractional integrals and Sobolev inequalities with applications to Schrödinger operators*, Comm. Partial Differential Equations, 10 (1985), pp. 1077-1116.

[8] C. L. FEFFERMAN, *The uncertainty principle*, Bull. Amer. Math. Soc., 9 (1983), pp. 129-206.

[9] M. SCHECHTER, *Hamiltonians for singular potentials*, Indiana Univ. Math. J., 5 (1972), pp. 483-503.

[10] A. DEVINATZ, *Schrödinger operators with singular potentials*, J. Operator Theory, 4 (1980), pp. 25-35.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

006

NONLOCAL VARIATIONAL PROBLEMS IN NONLINEAR ELECTROMAGNETO-ELASTOSTATICS*

ROBERT C. ROGERS†

Abstract. The effects of arbitrary applied electric and magnetic fields on unshielded, nonlinear, deformable, polarizable, magnetizable, nonconducting bodies are studied. Both monotone materials (dielectric, paramagnetic, etc.) and classical ferromagnetic materials are considered. The lack of shielding forces us to consider unknown fields outside of the material. This leads to nonlocal ("shape-dependent") effects. The work of Ball is extended ["Convexity conditions and existence theorems in nonlinear elasticity," *Arch. Rat. Mech. Anal.*, 63 (1977), pp. 337-403] to get an existence theory using direct methods of the calculus of variations.

Key words. electromagneto-elasticity, ferromagnetism, polyconvex materials

AMS (MOS) subject classifications. 73R05, 49A29, 73C60

1. Introduction.

Rogers and Antman [15] initiated a program to extend the modern existence theories for elastic materials (cf. [3], [8]) to materials that admit coupled elastic and electromagnetic effects. Such problems are of great practical interest (high current devices such as fusion reaction containment vessels and magnetic levitation trains and acoustic devices using piezoelectric materials are highly dependent on the coupling of mechanical and electromagnetic fields), and the nonlocal nature of the resultant equations is of intrinsic mathematical interest. The problems considered in [15] involved self-effects of conducting bodies; but for simplicity, electromagnetic boundary conditions that shielded the body from applied fields were employed. In this work, we consider an unshielded body and study the effects of applied electric and magnetic fields. We assume the body is nonconducting; but we allow it to be deformable, polarizable, and magnetizable. We study both monotone materials (dielectric, paramagnetic, etc.) and classical ferromagnetic materials.

Our problem is rather easy to formulate in spatial (Eulerian) coordinates (we do so in §2.3), but such a formulation is, in general, intractable. (For example, the boundary conditions are posed at points that depend on the deformation, a principal unknown.) A more useful formulation uses spatial coordinates for the resultant electric and magnetic fields E_r and H_r (which must be known both in the interior and exterior of the body) and material (Lagrangian) coordinates for the deformation $\hat{\mathbf{y}}$, deformation gradient $\mathbf{F}$, polarization $\mathbf{p}$, and magnetization $\mathbf{m}$. The variational form of this problem involves finding stationary points of an energy of the form

$$\begin{aligned}\tilde{\mathcal{E}}(\hat{\mathbf{y}}, E_r, \mathbf{p}, H_r, \mathbf{m}) &= \int_\Omega \mathcal{W}(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}))dv_x \\ &\quad - \int_\Omega \{[E_r(\hat{\mathbf{y}}(\mathbf{x})) + E_0(\hat{\mathbf{y}}(\mathbf{x}))] \cdot \mathbf{F}(\mathbf{x}) \cdot \mathbf{p}(\mathbf{x}) \\ &\quad\quad + [H_r(\hat{\mathbf{y}}(\mathbf{x})) + H_0(\hat{\mathbf{y}}(\mathbf{x}))] \cdot \mathbf{F}(\mathbf{x}) \cdot \mathbf{m}(\mathbf{x})\}dv_x \\ &\quad - \frac{1}{2}\int_{\mathbf{R}^3} |E_r(\mathbf{y})|^2 + |H_r(\mathbf{y})|^2 dv_y.\end{aligned}$$

* Received by the editors September 4, 1987; accepted for publication (in revised form) January 6, 1988.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The work of this author was completed at the Center for Mathematical Sciences and Department of Mathematics, University of Wisconsin-Madison, and was supported in part by the National Science Foundation under grants DMS–8521687 and DMS–8620303.

Here, $\mathcal{W}$ is the stored energy and E_0 and H_0 are the applied fields.

Unfortunately, this energy is neither coercive nor weakly lower-semicontinuous. We overcome this difficulty by replacing the spatial fields E_r and H_r with solution operators that depend on the *global* values of material fields:

$$\begin{aligned} E_r(\mathbf{y}) &= \hat{E}_r(\mathbf{p}(\cdot), \hat{\mathbf{y}}(\cdot); \mathbf{y}), \\ H_r(\mathbf{y}) &= \hat{H}_r(\mathbf{m}(\cdot), \hat{\mathbf{y}}(\cdot); \mathbf{y}). \end{aligned}$$

If $\mathbf{p}$, $\mathbf{m}$, and $\hat{\mathbf{y}}$ are sufficiently smooth these are simply the usual Coulomb integral operators. We show below that after the substitution of these nonlocal operators and some manipulation we get an energy of the form

$$\begin{aligned} \mathcal{E}(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) &= \int_\Omega \mathcal{W}(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}))dv_x \\ &\quad - \int_\Omega [E_0(\hat{\mathbf{y}}(\mathbf{x})) \cdot \mathbf{F}(\mathbf{x}) \cdot \mathbf{p}(\mathbf{x}) + H_0(\hat{\mathbf{y}}(\mathbf{x})) \cdot \mathbf{F}(\mathbf{x}) \cdot \mathbf{m}(\mathbf{x})]dv_x \\ &\quad + \frac{1}{2}\int_{\mathbf{R}^3} \{|\hat{E}_r(\mathbf{p}, \hat{\mathbf{y}}; \mathbf{y})|^2 + |\hat{H}_r(\mathbf{m}, \hat{\mathbf{y}}; \mathbf{y})|^2\}dv_y. \end{aligned}$$

Note that now the only unknowns are material fields; the spatial fields involved are either data or well-defined operators on material fields. Moreover, the new energy is coercive and (though this is not obvious) weakly lower-semicontinuous under appropriate hypotheses on $\mathcal{W}$. However, we have paid a price for these gains: While the first two integrals in the new energy involve *local* densities (e.g., at a particle $\mathbf{x}$, the stored energy density $\mathcal{W}$ depends only on the values of the fields at that particle), the final integral involves *nonlocal* energy densities (e.g., at every point $\mathbf{y}$, the electric field operator $\hat{E}_r$ depends on the *global* values of the fields $\mathbf{p}$ and $\hat{\mathbf{y}}$). The most novel difficulties in our existence theory involve these nonlocal terms. In particular, we must examine the weak convergence properties of the solution operators $\hat{E}$ and $\hat{H}$ under composite limits of sequences of deformations, polarizations, and magnetizations.

Variational problems with nonlocal energy densities have been considered before. Notably, Auchmuty and Beals [2] gave existence and regularity theorems for some model problems from astrophysics in which self-gravitation plays an important role.

The remainder of this paper is organized as follows: In §2 we present several mathematical formulations of our physical problem, culminating in the nonlocal variational problem that we eventually solve. In §3 we give some results on how the electric and magnetic fields react under simultaneous limits of deformations, polarizations, and magnetizations. In §4 we prove existence results. Finally, in §5 we offer some concluding observations.

2. Formulation of the problem. In this section we formulate the problem of a deformable, polarizable, magnetizable, nonconducting body subjected to arbitrary applied static electromagnetic fields.

2.1. Kinematics. We consider a body whose reference configuration is given by a bounded set Ω in $\mathbf{R}^3$ with typical material particle $\mathbf{x}$ and Lipschitz continuous boundary. We refer to fields with independent variable $\mathbf{x}$ as *material* or *Lagrangian*. The familiar differential operators Grad, Div, and Curl operate on such fields. The deformation of the body Ω is given by the map

$$\Omega \ni \mathbf{x} \mapsto \hat{\mathbf{y}}(\mathbf{x}) \in \mathbf{R}^3. \tag{2.1}$$

The operators grad, div, and curl operate on fields in *spatial* or *Eulerian* coordinates (fields with typical independent variable $\mathbf{y}$). The *deformation gradient* $\mathbf{F}$ and the *right Cauchy-Green deformation tensor* $\mathbf{C}$ are given by

$$\mathbf{F}(\mathbf{x}) = \operatorname{Grad} \hat{\mathbf{y}}(\mathbf{x})^*, \qquad \mathbf{C}(\mathbf{x}) = \mathbf{F}^*(\mathbf{x}) \cdot \mathbf{F}(\mathbf{x}). \tag{2.2}$$

Here the star $*$ indicates *transpose*, and $\mathbf{A} \cdot \mathbf{B}$ indicates the *product* of two tensors. To ensure that the orientation of the material is preserved under deformation we require

$$\det \mathbf{F} > 0. \tag{2.3}$$

To ensure that there is no interpenetration of the material we require (cf. Ciarlet and Nečas [8]):

$$\int_\Omega \det \mathbf{F}(\mathbf{x}) dv_x \leq |\hat{\mathbf{y}}(\Omega)|, \tag{2.4}$$

where $|S|$ is the three-dimensional Lebesgue measure of the set S.

2.2. Applied fields. We assume that the total electric and magnetic fields E and H can be decomposed as

$$E = E_0 + E_r \quad \text{and} \quad H = H_0 + H_r,$$

where E_0 and H_0 are the respective *applied electric and magnetic fields* (fields that are generated by sources not connected to the body) and E_r and H_r are the respective *resultant electric and magnetic fields* (fields generated by the polarization and magnetization of the body).

We think of the applied fields as data and assume only that they satisfy

$$E_0, H_0 \in \{\mathbf{v} \in L^2(\mathbf{R}^3) | \mathbf{v} = \operatorname{grad} \phi, \phi \in H^1_{\text{loc}}(\mathbf{R}^3)\}. \tag{2.5}$$

Here $H^s = W^{s,2}$ where $W^{s,p}$ is the usual Sobolev space of functions with generalized derivatives of order up to s in L^p (cf. Adams [1]), and $H^1_{\text{loc}}(\mathbf{R}^3)$ is the space of functions ϕ such that $\phi\psi \in H^1(\mathbf{R}^3)$ for any $\psi \in C_0^\infty(\mathbf{R}^3)$. Our assumption implies that the applied fields are fixed and are not changed by the resultant fields. We think of the sources of these fields as being outside the deformed body, though this is not implied by (2.5).

2.3. Spatial formulation. In this formulation our principal unknowns are the deformation $\hat{\mathbf{y}}$ and the resultant electric and magnetic fields E_r and H_r defined above, the *Cauchy stress* $\mathbf{T}$, the *spatial polarization* P, and the *spatial magnetization* M. We require that the support of $\mathbf{T}$, P, and M be contained in the image of the deformed body $\hat{\mathbf{y}}(\Omega)$.

We assume that the Cauchy stress $\mathbf{T}$ can be decomposed[1] in the following way:

$$\mathbf{T} = \mathbf{T}_M + \mathbf{T}_{ms}, \tag{2.6}$$

[1] This decomposition is not unique. Hutter and van de Ven [12] discuss various possible formulations of electromagnetism in deformable media and subsequent variations in the form of the Maxwell stress tensor. In the static case, all of the formulations they consider are shown to be equivalent under appropriate choices of constitutive equations. Our constitutive assumptions are compatible with any of these choices.

where $\mathbf{T}_M$ is the *mechanical stress* and $\mathbf{T}_{ms}$ is the *Maxwell stress*, which is given by

$$\mathbf{T}_{ms} = EE + EP + HH + HM - \frac{1}{2}[E \cdot (E+P) + H \cdot (H+M)]\mathbf{I}. \tag{2.7}$$

Here EE, EP, etc. are dyadic (tensor) products and $\mathbf{I}$ is the identity tensor.

We also assume that the *electromagnetic body couple* $\mathbf{L}$ is given by

$$\mathbf{L} = EP - PE + HM - MH. \tag{2.8}$$

(See Rogers and Antman [15] for a discussion of the absorption of electromagnetic body forces into a generalized stress tensor and the subsequent form of the body couple.)

We now state a local, spatial formulation of our physical problem. For the present, we assume that all of our functions are smooth enough for the equations in which they appear to hold in a classical sense.

Problem 2.1. Given functions $\mathbf{y}_0$, $\mathbf{s}_0$, E_0, and H_0, we seek $\hat{\mathbf{y}}$, $\mathbf{T}$, E_r, H_r, P, and M such that:

1. The deformation $\hat{\mathbf{y}}$ satisfies (2.3) and (2.4). At points $\mathbf{y} \in \hat{\mathbf{y}}(\Omega)$ inside the deformed body the *balance of linear momentum*

$$\operatorname{div} \mathbf{T} = \mathbf{0}, \tag{2.9}$$

the *balance of torque*

$$\mathbf{L} = \mathbf{T} - \mathbf{T}^*, \tag{2.10}$$

and the *static version of Maxwell's equations*

$$\operatorname{curl} E_r(\mathbf{y}) = \mathbf{0}, \tag{2.11}$$

$$\operatorname{div} E_r(\mathbf{y}) = -\operatorname{div} P(\mathbf{y}), \tag{2.12}$$

$$\operatorname{curl} H_r(\mathbf{y}) = \mathbf{0}, \tag{2.13}$$

$$\operatorname{div} H_r(\mathbf{y}) = -\operatorname{div} M(\mathbf{y}), \tag{2.14}$$

are satisfied.

2. We assume that the boundary of the reference configuration $\partial\Omega$ can be represented as the disjoint union of two sets with Lipschitz boundary, S_1 and S_2. On S_1 we prescribe Dirichlet boundary conditions

$$\hat{\mathbf{y}}(\mathbf{x}) = \mathbf{y}_0(\mathbf{x}), \quad \mathbf{x} \in S_1, \tag{2.15}$$

and on the deformed image of S_2 we prescribe dead-load boundary conditions

$$\mathbf{T}(\hat{\mathbf{y}}(\mathbf{x})) \cdot \hat{\mathbf{n}}(\hat{\mathbf{y}}(\mathbf{x})) = \det \mathbf{F}(\mathbf{x})\mathbf{s}_0(\mathbf{x}), \quad \mathbf{x} \in S_2. \tag{2.16}$$

Here $\hat{\mathbf{n}}$ is the unit outward normal to $\partial\hat{\mathbf{y}}(\Omega)$. On the entire surface $\partial\hat{\mathbf{y}}(\Omega)$ we require

$$[|E_r(\mathbf{y})|] \times \hat{\mathbf{n}}(\mathbf{y}) = 0, \tag{2.17}$$

$$[|E_r(\mathbf{y})|] \cdot \hat{\mathbf{n}}(\mathbf{y}) = -[|P(\mathbf{y})|] \cdot \hat{\mathbf{n}}(\mathbf{y}), \tag{2.18}$$

$$[|H_r(\mathbf{y})|] \times \hat{\mathbf{n}}(\mathbf{y}) = 0, \tag{2.19}$$

$$[|H_r(\mathbf{y})|] \cdot \hat{\mathbf{n}}(\mathbf{y}) = -[|M(\mathbf{y})|] \cdot \hat{\mathbf{n}}(\mathbf{y}). \tag{2.20}$$

Here $[|f|]$ indicates the jump in the field f in the direction of $\hat{\mathbf{n}}$.

3. At points $\mathbf{y} \in \mathbf{R}^3 \setminus \hat{\mathbf{y}}(\bar{\Omega})$ in the exterior of the body we require

$$\text{div } E = 0, \tag{2.21}$$

$$\text{curl } E = 0, \tag{2.22}$$

$$\text{div } H = 0, \tag{2.23}$$

$$\text{curl } H = 0. \tag{2.24}$$

4. We also require that the resultant electric and magnetic fields be *regular at infinity*

$$|\mathbf{y}|^2|E_r(\mathbf{y})| + |\mathbf{y}|^2|H_r(\mathbf{y})| = O(1) \text{ as } |\mathbf{y}| \to \infty. \tag{2.25}$$

While this statement of the problem is clear, it presents several impediments to a solution: One set of difficulties is common to all elasticity problems posed entirely in spatial coordinates (e.g., constitutive theory for nonhomogeneous media is difficult to formulate and the support of the unknowns is itself an unknown). In the next section we present a formulation that overcomes these problems by introducing material coordinates. A second set of troubles arises from the addition of electromagnetic effects and the lack of shielding. This forces us to solve for the electric and magnetic fields in the exterior of the body with jump conditions at the unknown points of the deformed boundary. The solution of this problem must wait until §2.7 where we introduce solution operators that give the electric and magnetic fields in terms of material fields.

2.4. Mixed formulation. The *referential* or *(First) Piola-Kirchhoff stress* $\mathbf{S}$ at a material particle $\mathbf{x} \in \Omega$ is defined by

$$\mathbf{S}(\mathbf{x})^* = \det \mathbf{F}(\mathbf{x})\mathbf{F}^{-1}(\mathbf{x}) \cdot \mathbf{T}(\hat{\mathbf{y}}(\mathbf{x}))^*. \tag{2.26}$$

The Piola-Kirchhoff versions of the mechanical and Maxwell portions of the stress ($\mathbf{S}_M$ and $\mathbf{S}_{ms}$) have an analogous relation to their spatial counterparts. We also introduce the *material versions of the polarization and magnetization*:

$$\mathbf{p}(\mathbf{x}) = \det \mathbf{F}(\mathbf{x})\ \mathbf{F}^{-1}(\mathbf{x}) \cdot P(\hat{\mathbf{y}}(\mathbf{x})), \tag{2.27}$$

$$\mathbf{m}(\mathbf{x}) = \det \mathbf{F}(\mathbf{x})\ \mathbf{F}^{-1}(\mathbf{x}) \cdot M(\hat{\mathbf{y}}(\mathbf{x})), \tag{2.28}$$

and the *material version of the electromagnetic body couple*:

$$\tilde{\mathbf{L}}(\mathbf{x}) = (\det \mathbf{F}(\mathbf{x}))\mathbf{L}(\hat{\mathbf{y}}(\mathbf{x})). \tag{2.29}$$

Since we need to consider the electric and magnetic fields at points in the exterior of the body, we are primarily interested in their spatial versions, but at points in the interior of the body we can introduce *material versions of the electric and magnetic fields*:

$$\mathbf{e}(\mathbf{x}) = E(\hat{\mathbf{y}}(\mathbf{x})) \cdot \mathbf{F}(\mathbf{x}), \tag{2.30}$$

$$\mathbf{h}(\mathbf{x}) = H(\hat{\mathbf{y}}(\mathbf{x})) \cdot \mathbf{F}(\mathbf{x}). \tag{2.31}$$

Material versions of the applied and resultant fields ($\mathbf{e}_0$ and $\mathbf{e}_r$) are defined in an analogous way.

Three pieces of our data remain unchanged: The applied electric and magnetic fields E_0 and H_0 are still best given in spatial coordinates, and the Dirichlet data $\mathbf{y}_0$ was always given as a material field. However, we will replace the live-load traction boundary condition and the spatial data $\mathbf{t}_0$ with a dead-load condition and material data $\mathbf{s}_0$. As we indicated above, we consider the dead-load condition our primary problem and merely introduced the live-load condition to present a similar problem with an easy exposition. With this change, we state Problem 2.1 in terms of the material coordinates introduced above.

Problem 2.2. Given functions $\mathbf{y}_0$, $\mathbf{s}_0$, E_0, and H_0, we seek $\hat{\mathbf{y}}$, $\mathbf{S}$, E_r, H_r, $\mathbf{p}$, and $\mathbf{m}$ such that:

1. The deformation $\hat{\mathbf{y}}$ satisfies (2.3) and (2.4). At material particles $\mathbf{x} \in \Omega$ we have

$$\text{Div } \mathbf{S} = \mathbf{0}, \tag{2.32}$$

$$\tilde{\mathbf{L}} = \mathbf{S} \cdot \mathbf{F}^* - \mathbf{F} \cdot \mathbf{S}^*, \quad \text{and} \tag{2.33}$$

$$\text{Curl } \mathbf{e}_r(\mathbf{x}) = \mathbf{0}, \tag{2.34}$$

$$\text{Div } (\det \mathbf{F}(\mathbf{x})\mathbf{C}^{-1}(\mathbf{x}) \cdot \mathbf{e}_r(\mathbf{x})) = -\text{Div } \mathbf{p}(\mathbf{x}), \tag{2.35}$$

$$\text{Curl } \mathbf{h}_r(\mathbf{x}) = \mathbf{0}, \tag{2.36}$$

$$\text{Div } (\det \mathbf{F}(\mathbf{x})\mathbf{C}^{-1}(\mathbf{x}) \cdot \mathbf{h}_r(\mathbf{x})) = -\text{Div } \mathbf{m}(\mathbf{x}). \tag{2.37}$$

2. The Dirichlet conditions

$$\hat{\mathbf{y}}(\mathbf{x}) = \mathbf{y}_0(\mathbf{x}) \quad \forall \mathbf{x} \in S_1 \tag{2.38}$$

are satisfied, the dead-load conditions

$$\mathbf{S}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) = \mathbf{s}_0(\mathbf{x}) \quad \forall \mathbf{x} \in S_2 \tag{2.39}$$

are satisfied, and for every $\mathbf{x} \in \partial\Omega$ the jump conditions

$$[|E_r(\hat{\mathbf{y}}(\mathbf{x}))|] \cdot \mathbf{F}(\mathbf{x}) \times \cdot \mathbf{n}(\mathbf{x}) = 0, \tag{2.40}$$

$$\det \mathbf{F}(\mathbf{x})\mathbf{C}(\mathbf{x})^{-1} \cdot [|E_r(\hat{\mathbf{y}}(\mathbf{x}))|] \cdot \mathbf{n}(\mathbf{x}) = \mathbf{p}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}), \tag{2.41}$$

$$[|H_r(\hat{\mathbf{y}}(\mathbf{x}))|] \cdot \mathbf{F}(\mathbf{x}) \times \mathbf{n}(\mathbf{x}) = 0, \tag{2.42}$$

$$\det \mathbf{F}(\mathbf{x})\mathbf{C}(\mathbf{x})^{-1} \cdot [|H_r(\hat{\mathbf{y}}(\mathbf{x}))|] \cdot \mathbf{n}(\mathbf{x}) = \mathbf{m}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}), \tag{2.43}$$

are satisfied. Here $\mathbf{n}$ is the unit outward normal to $\partial\Omega$.

3. In the exterior of the body E_r and H_r satisfy (2.21)–(2.25) just as they did in Problem 2.1.

Of course, Problems 2.1 and 2.2 are both underdetermined. We introduce constitutive equations to remedy this.

2.5. Constitutive equations. Our constitutive equations for electromagnetoelastic materials are fully coupled and nonlinear. (Thus, we allow for such effects as electrostriction and magnetostriction.) In the variational problems we consider, it is

convenient to adopt $\mathbf{F}$, $\mathbf{p}$, and $\mathbf{m}$ as independent constitutive variables and $\mathbf{S}$, $\mathbf{e}$, and $\mathbf{h}$ as dependent constitutive variables.[2]

We first direct our attention to materials that exhibit monotone electromagnetic effects. These include many common materials such as dielectrics, certain pyroelectrics, paramagnetics, and even (at least for approximate constitutive equations) soft ferromagnetics. Our constitutive equations have the general form:

$$\begin{aligned} \mathbf{S}_M(\mathbf{x}) &= \tilde{\mathbf{S}}_M(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}), \mathbf{x}), \\ \mathbf{e}(\mathbf{x}) &= \tilde{\mathbf{e}}(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}), \mathbf{x}), \\ \mathbf{h}(\mathbf{x}) &= \tilde{\mathbf{h}}(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}), \mathbf{x}), \end{aligned} \tag{2.47}$$

but we consider only materials for which there is a *stored energy function* $\mathcal{W}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x})$, continuously differentiable in $\mathbf{F}$, $\mathbf{p}$, and $\mathbf{m}$ and measurable in $\mathbf{x}$ for all values of the remaining arguments, such that

$$\begin{aligned} \tilde{\mathbf{S}}_M(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}) &= \frac{\partial \mathcal{W}}{\partial \mathbf{F}}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}), \\ \tilde{\mathbf{e}}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}) &= \frac{\partial \mathcal{W}}{\partial \mathbf{p}}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}), \\ \tilde{\mathbf{h}}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}) &= \frac{\partial \mathcal{W}}{\partial \mathbf{m}}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}). \end{aligned} \tag{2.48}$$

In the existence theorems of §4 we state our convexity and growth hypotheses on $\mathcal{W}$.

We also consider ferromagnetics: a type of nonmonotone material. For this material we drop consideration of electric effects; we assume the body does not polarize and that the solutions of the electric field equations are independent of solutions of the balance laws of elasticity and magnetism. Following the classical model of ferromagnetism (cf. Landau and Lifshitz [14] and Brown [5]), we consider a stored energy function

$$\mathcal{V}(\mathbf{F}, \mathbf{m}, \mathbf{x}) \tag{2.49}$$

[2] In other problems it is convenient to take $\mathbf{F}$, $\mathbf{e}$, and $\mathbf{h}$ as independent variables and $\mathbf{S}$, $\mathbf{p}$, and $\mathbf{m}$ as dependent variables:

$$\begin{aligned} \mathbf{S}(\mathbf{x}) &= \check{\mathbf{S}}(\mathbf{F}(\mathbf{x}), \mathbf{e}(\mathbf{x}), \mathbf{h}(\mathbf{x}), \mathbf{x}), \\ \mathbf{p}(\mathbf{x}) &= \check{\mathbf{p}}(\mathbf{F}(\mathbf{x}), \mathbf{e}(\mathbf{x}), \mathbf{h}(\mathbf{x}), \mathbf{x}), \\ \mathbf{m}(\mathbf{x}) &= \check{\mathbf{m}}(\mathbf{F}(\mathbf{x}), \mathbf{e}(\mathbf{x}), \mathbf{h}(\mathbf{x}), \mathbf{x}). \end{aligned} \tag{2.44}$$

Such a choice can be shown to be equivalent to the one chosen in this work under appropriate monotonicity and growth hypotheses. For instance, we could adopt a monotonicity assumption such as the *restricted strong ellipticity condition*:

$$\begin{aligned} 0 \;<\; & \mathbf{ab} : \frac{\partial \check{\mathbf{S}}}{\partial \mathbf{F}} : \mathbf{ab} + \mathbf{ab} : \frac{\partial \check{\mathbf{S}}}{\partial \mathbf{e}} \cdot \mathbf{u} + \mathbf{ab} : \frac{\partial \check{\mathbf{S}}}{\partial \mathbf{h}} \cdot \mathbf{v} \\ & + \mathbf{u} \cdot \frac{\partial \check{\mathbf{p}}}{\partial \mathbf{F}} : \mathbf{ab} + \mathbf{u} \cdot \frac{\partial \check{\mathbf{p}}}{\partial \mathbf{e}} \cdot \mathbf{u} + \mathbf{u} \cdot \frac{\partial \check{\mathbf{p}}}{\partial \mathbf{h}} \cdot \mathbf{v} \\ & + \mathbf{v} \cdot \frac{\partial \check{\mathbf{m}}}{\partial \mathbf{F}} : \mathbf{ab} + \mathbf{v} \cdot \frac{\partial \check{\mathbf{m}}}{\partial \mathbf{e}} \cdot \mathbf{u} + \mathbf{v} \cdot \frac{\partial \check{\mathbf{m}}}{\partial \mathbf{h}} \cdot \mathbf{v} \end{aligned} \tag{2.45}$$

for all $(\mathbf{ab}, \mathbf{u}, \mathbf{v}) \neq (\mathbf{0}, \mathbf{0}, \mathbf{0})$ ($\mathbf{ab}$ is a dyad, a typical tensor of rank one) and a growth condition such as

$$[\check{\mathbf{p}}(\mathbf{F}, \mathbf{e}, \mathbf{h}, \mathbf{x}) \cdot \mathbf{e} + \check{\mathbf{m}}(\mathbf{F}, \mathbf{e}, \mathbf{h}, \mathbf{x}) \cdot \mathbf{h}]/[|\mathbf{e}| + |\mathbf{h}|] \to \infty \text{ as } [|\mathbf{e}| + |\mathbf{h}|] \to \infty. \tag{2.46}$$

Equations (2.45) and (2.46) imply that for every fixed $\mathbf{F}$ and $\mathbf{x}$, the map $(\mathbf{e}, \mathbf{h}) \mapsto (\check{\mathbf{p}}, \check{\mathbf{m}})$ can be globally inverted, so that constitutive equations (2.44) are equivalent to (2.47).

for which $\mathcal{V}(\mathbf{F}, \cdot, \mathbf{x})$ is *nonconvex.* We also include in the total energy (cf. (2.74) below) an *exchange energy* term of the form

$$\mathcal{X}(\mathbf{F}, \text{Grad } \mathbf{m}) \tag{2.50}$$

for which we make appropriate convexity assumptions below. The most commonly used term is of the form $|\text{Grad } \mathbf{m}|^2$. Once again, we reserve more detailed continuity and growth hypotheses until the statement of our existence theorems in §4.

2.6. Mixed variational formulation with local densities. In this section we present a variational formulation for monotone materials. We postpone consideration of ferromagnetic materials until the final, nonlocal formulation.

The usual variational form of Problem 2.2 is obtained by defining an energy depending on *all* of the unknown electromagnetic fields:

$$\begin{aligned}\tilde{\mathcal{E}}(\hat{\mathbf{y}}, E_r, \mathbf{p}, H_r, \mathbf{m}) &= \int_\Omega \mathcal{W}(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}), \mathbf{x}) dv_x \\ &\quad - \int_\Omega \{[E_r(\hat{\mathbf{y}}(\mathbf{x})) + E_0(\hat{\mathbf{y}}(\mathbf{x}))] \cdot \mathbf{F}(\mathbf{x}) \cdot \mathbf{p}(\mathbf{x}) \\ &\quad + [H_r(\hat{\mathbf{y}}(\mathbf{x})) + H_0(\hat{\mathbf{y}}(\mathbf{x}))] \cdot \mathbf{F}(\mathbf{x}) \cdot \mathbf{m}(\mathbf{x})\} dv_x \\ &\quad - \frac{1}{2} \int_{\mathbf{R}^3} |E_r(\mathbf{y})|^2 + |H_r(\mathbf{y})|^2 dv_y.\end{aligned} \tag{2.51}$$

Problem 2.3. Given functions $\mathbf{y}_0$, $\mathbf{s}_0$, E_0, and H_0, we seek $\hat{\mathbf{y}}$, E_r, H_r, $\mathbf{p}$, and $\mathbf{m}$ such that:

1. The deformation $\hat{\mathbf{y}}$ satisfies (2.3), (2.4), and (2.38).
2. The fields E_r and H_r satisfy (2.11) and (2.13), respectively.
3. The variational equation

$$\delta \tilde{\mathcal{E}}(\hat{\mathbf{y}}, E_r, \mathbf{p}, H_r, \mathbf{m}) = \int_{S_2} \mathbf{s}_0(\mathbf{x}) \cdot \delta \hat{\mathbf{y}}(\mathbf{x}) da_x, \tag{2.52}$$

is satisfied.

It follows from the work of Toupin [17] and Tiersten [16] that sufficiently regular solutions of Problem 2.3 are solutions of Problem 2.2 as well.

Two difficulties to finding solutions of Problem 2.3 are immediately apparent. First, solutions of (2.52) are not extrema of (2.51); we must simultaneously minimize $\tilde{\mathcal{E}}$ with respect to $\hat{\mathbf{y}}$, $\mathbf{p}$, and $\mathbf{m}$ and maximize with respect to E_r and H_r. Second, the comments at the end of §2.3 still hold: the use of spatial coordinates for the fields E_r and H_r give us trouble; we must optimize these fields without knowing in advance the points at which they must jump, much less the magnitude or direction of the jump. To eliminate these difficulties, we define, in the section below, solution operators that give the electric and magnetic fields in terms of the material fields $\hat{\mathbf{y}}$, $\mathbf{p}$, and $\mathbf{m}$.

2.7. Solution operators. In defining our first solution operator we think of the spatial fields P and M as being prescribed functions in $L^2(\mathbf{R}^3)$ with support in the deformed body $\hat{\mathbf{y}}(\Omega)$, and we solve for the electric and magnetic fields they generate. Note that we can do this without reference to the deformation.

PROPOSITION 2.1. *Let*

$$\mathcal{A} \equiv \{V \in L^2(\mathbf{R}^3) | \text{curl } V = \mathbf{0} \quad \text{in } H^{-1}(\mathbf{R}^3)\}. \tag{2.53}$$

For every pair of functions $(P, M) \in L^2(\hat{\mathbf{y}}(\Omega))$ *there exists a unique pair* $(E_r, H_r) \in \mathcal{A}$*:*

$$\begin{aligned} E_r(\mathbf{y}) &= \check{E}_r(P;\mathbf{y}), \\ H_r(\mathbf{y}) &= \check{H}_r(M;\mathbf{y}) \end{aligned}$$

that satisfies

$$\int_{\mathbf{R}^3} (E_r \cdot E^\sharp + H_r \cdot H^\sharp) dv = -\int_{\hat{\mathbf{y}}(\Omega)} (P \cdot E^\sharp + M \cdot H^\sharp) dv \quad \forall\, E^\sharp, H^\sharp \in \mathcal{A}. \tag{2.54}$$

Furthermore,

$$\|\check{E}_r(P;\cdot)\|_{L^2(\mathbf{R}^3)} \le C\|P\|_{L^2(\hat{\mathbf{y}}(\Omega))}, \quad \textit{and} \tag{2.55}$$

$$\|\check{H}_r(M;\cdot)\|_{L^2(\mathbf{R}^3)} \le C\|M\|_{L^2(\hat{\mathbf{y}}(\Omega))}. \tag{2.56}$$

If P *and* M *are in* $H^1(\hat{\mathbf{y}}(\Omega))$*, then the solution operators are given by the following integral equations:*

$$\begin{aligned} E_r(\mathbf{y}) &= \check{E}_r(P;\mathbf{y}) \\ &\equiv \tfrac{1}{4\pi}\left[\int_{\hat{\mathbf{y}}(\Omega)} \frac{-\operatorname{div} P(\mathbf{y}')(\mathbf{y}-\mathbf{y}')}{|\mathbf{y}-\mathbf{y}'|^3} dv_{y'} + \int_{\partial\hat{\mathbf{y}}(\Omega)} \frac{P(\mathbf{y}')\cdot\hat{\mathbf{n}}(\mathbf{y}')(\mathbf{y}-\mathbf{y}')}{|\mathbf{y}-\mathbf{y}'|^3} da_{y'}\right]. \end{aligned} \tag{2.57}$$

$$\begin{aligned} H_r(\mathbf{y}) &= \check{H}_r(M;\mathbf{y}) \\ &\equiv \tfrac{1}{4\pi}\left[\int_{\hat{\mathbf{y}}(\Omega)} \frac{-\operatorname{div} M(\mathbf{y}')(\mathbf{y}-\mathbf{y}')}{|\mathbf{y}-\mathbf{y}'|^3} dv_{y'} + \int_{\partial\hat{\mathbf{y}}(\Omega)} \frac{M(\mathbf{y}')\cdot\hat{\mathbf{n}}(\mathbf{y}')(\mathbf{y}-\mathbf{y}')}{|\mathbf{y}-\mathbf{y}'|^3} da_{y'}\right]. \end{aligned} \tag{2.58}$$

Here, $\hat{\mathbf{n}}$ *is the unit outward normal to* $\partial\hat{\mathbf{y}}(\Omega)$*.*

If P *and* M *are in* $C^1(\hat{\mathbf{y}}(\bar{\Omega}))$ *then* E_r *and* H_r *satisfy* (2.11)–(2.14) *in the interior and exterior of the body,* (2.17)–(2.20) *at the boundary of the body, and are regular at infinity.*

This result follows from the Lax-Milgram lemma and standard results of potential theory (cf., e.g., Kellogg [13]).

We now define solution operators that tie E_r and H_r directly to the material fields $\hat{\mathbf{y}}$, $\mathbf{p}$, and $\mathbf{m}$. Let $(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m})$ be prescribed functions in the set

$$\mathcal{B} \equiv \left\{ (\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \;\middle|\; \begin{array}{l} \hat{\mathbf{y}} \in W^{1,p}(\Omega), p \ge 2; \text{(2.15) holds in the sense of trace;} \\ \mathbf{F}^\times \in L^q(\Omega), q \ge (p-1)/p; \;\; \det\mathbf{F} \in L^r(\Omega), r > 1; \\ \det\mathbf{F} > 0 \text{ a.e. in } \Omega; \;\; \int_\Omega \det\mathbf{F} dv \le |\hat{\mathbf{y}}(\Omega)|; \\ \dfrac{\mathbf{F}\cdot\mathbf{p}}{\sqrt{\det\mathbf{F}}}, \dfrac{\mathbf{F}\cdot\mathbf{m}}{\sqrt{\det\mathbf{F}}} \in L^2(\Omega). \end{array} \right\} \tag{2.59}$$

Here $\mathbf{F}^\times$ indicates the formal adjoint or cofactor matrix of $\mathbf{F}$. Before continuing we observe that for any $(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \in \mathcal{B}$ we have:

1. Ciarlet and Nečas [8] imply that the map $\hat{\mathbf{y}} : \Omega \mapsto \mathbf{R}^3$ is invertible on $\hat{\mathbf{y}}(\Omega)$. We denote the inverse by $\hat{\mathbf{x}}$.

2. The composite functions

$$\tilde{P}(\hat{\mathbf{x}}, \mathbf{p}; \mathbf{y}) \equiv \begin{cases} [\det\mathbf{F}(\hat{\mathbf{x}}(\mathbf{y}))]^{-1}\mathbf{F}(\hat{\mathbf{x}}(\mathbf{y}))\cdot\mathbf{p}(\hat{\mathbf{x}}(\mathbf{y})), & \mathbf{y} \in \hat{\mathbf{y}}(\Omega), \\ \mathbf{0} & \text{elsewhere} \end{cases} \tag{2.60}$$

and

$$\tilde{M}(\hat{\mathbf{x}}, \mathbf{m}; \mathbf{y}) \equiv \begin{cases} [\det\mathbf{F}(\hat{\mathbf{x}}(\mathbf{y}))]^{-1}\mathbf{F}(\hat{\mathbf{x}}(\mathbf{y}))\cdot\mathbf{m}(\hat{\mathbf{x}}(\mathbf{y})), & \mathbf{y} \in \hat{\mathbf{y}}(\Omega), \\ \mathbf{0} & \text{elsewhere} \end{cases} \tag{2.61}$$

are in $L^2(\mathbf{R}^3)$.

3. Hölder's inequality implies that the functions $\mathbf{F}\cdot\mathbf{p}$ and $\mathbf{F}\cdot\mathbf{m}$ are in $L^s(\Omega)$ with $s=\frac{2r}{r+1}$.

We also observe that without further assumptions on p, q, and r we cannot ensure that $\mathbf{p}$ and $\mathbf{m}$ are in an L^p space with $p\geq 1$.

Now, as in the spatial case, we can define solution operators that describe the electric and magnetic fields as being generated by material fields in the set $\mathcal{B}$.

PROPOSITION 2.2. *For every triple of functions* $(\hat{\mathbf{y}},\mathbf{p},\mathbf{m})\in\mathcal{B}$ *there exists a unique pair* $(E_r,H_r)\in\mathcal{A}$*:*

$$\begin{aligned} E_r(\mathbf{y}) &= \hat{E}_r(\hat{\mathbf{y}},\mathbf{p};\mathbf{y}),\\ H_r(\mathbf{y}) &= \hat{H}_r(\hat{\mathbf{y}},\mathbf{m};\mathbf{y}) \end{aligned}$$

that satisfies

$$\begin{aligned} &\int_{\mathbf{R}^3} E_r(\mathbf{y})\cdot E^\sharp(\mathbf{y})+H_r(\mathbf{y})\cdot H^\sharp(\mathbf{y})\,dv_y\\ &\qquad=-\int_\Omega[\mathbf{F}(\mathbf{x})\cdot\mathbf{p}(\mathbf{x})\cdot E^\sharp(\hat{\mathbf{y}}(\mathbf{x}))+\mathbf{F}(\mathbf{x})\cdot\mathbf{m}(\mathbf{x})\cdot H^\sharp(\hat{\mathbf{y}}(\mathbf{x}))]\,dv_x \end{aligned} \tag{2.62}$$

for every $(E^\sharp,H^\sharp)\in\mathcal{A}$. *Furthermore,*

$$\|\hat{E}_r(\hat{\mathbf{y}},\mathbf{p};\cdot)\|_{L^2(\mathbf{R}^3)}\leq C\|\mathbf{F}\cdot\mathbf{p}/\sqrt{\det\mathbf{F}}\|_{L^2(\Omega)},\quad \textit{and} \tag{2.63}$$

$$\|\hat{H}_r(\hat{\mathbf{y}},\mathbf{m};\cdot)\|_{L^2(\mathbf{R}^3)}\leq C\|\mathbf{F}\cdot\mathbf{m}/\sqrt{\det\mathbf{F}}\|_{L^2(\Omega)}. \tag{2.64}$$

The material versions of these solution operators can then be defined by

$$\mathbf{e}_r(\mathbf{x})=\hat{\mathbf{e}}_r(\hat{\mathbf{y}},\mathbf{p};\mathbf{x})\equiv\hat{E}_r(\hat{\mathbf{y}},\mathbf{p};\hat{\mathbf{y}}(\mathbf{x}))\cdot\mathbf{F}(\mathbf{x}), \tag{2.65}$$

$$\mathbf{h}_r(\mathbf{x})=\hat{\mathbf{h}}_r(\hat{\mathbf{y}},\mathbf{m};\mathbf{x})\equiv\hat{H}_r(\hat{\mathbf{y}},\mathbf{m};\hat{\mathbf{y}}(\mathbf{x}))\cdot\mathbf{F}(\mathbf{x}). \tag{2.66}$$

As before, the proof of this result follows directly from the Lax-Milgram lemma.

Note that the operators $\check{E}_r$ and $\check{H}_r$ (which operate on spatial fields) and the operators $\hat{E}_r$ and $\hat{H}_r$ (which operate on material fields) are related by

$$\hat{E}_r(\hat{\mathbf{y}},\mathbf{p};\mathbf{y}) = \check{E}_r(\tilde{P}(\hat{\mathbf{x}},\mathbf{p};\cdot);\mathbf{y}), \tag{2.67}$$

$$\hat{H}_r(\hat{\mathbf{y}},\mathbf{m};\mathbf{y}) = \check{H}_r(\tilde{M}(\hat{\mathbf{x}},\mathbf{m};\cdot);\mathbf{y}), \tag{2.68}$$

where $\tilde{P}$ and $\tilde{M}$ are defined in (2.60) and (2.61).

2.8. Energy with nonlocal densities. Given any material functions $(\hat{\mathbf{y}},\mathbf{p},\mathbf{m})\in\mathcal{B}$ we can compose the solution operators $\hat{E}_r$ and $\hat{H}_r$ with $\tilde{\mathcal{E}}$ and use the relation

$$\begin{aligned} &\int_{\mathbf{R}^3}|\hat{E}_r(\mathbf{p},\hat{\mathbf{y}};\mathbf{y})|^2+|\hat{H}_r(\mathbf{m},\hat{\mathbf{y}};\mathbf{y})|^2\,dv_y\\ &\qquad=-\int_\Omega\hat{\mathbf{e}}_r(\mathbf{p},\hat{\mathbf{y}};\mathbf{x})\cdot\mathbf{p}(\mathbf{x})+\hat{\mathbf{h}}_r(\mathbf{m},\hat{\mathbf{y}};\mathbf{x})\cdot\mathbf{m}(\mathbf{x})\,dv_x \end{aligned} \tag{2.69}$$

(which is simply (2.62) with $E^\sharp=\hat{E}_r$ and $H^\sharp=\hat{H}_r$) to get the following representation for the energy:

$$\tilde{\mathcal{E}}(\hat{\mathbf{y}},\hat{E}_r(\mathbf{p},\hat{\mathbf{y}};\cdot),\mathbf{p},\hat{H}_r(\mathbf{m},\hat{\mathbf{y}};\cdot),\mathbf{m})=\mathcal{L}(\hat{\mathbf{y}},\mathbf{p},\mathbf{m})+\mathcal{G}(\hat{\mathbf{y}},\mathbf{p},\mathbf{m}), \tag{2.70}$$

where

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \equiv \int_\Omega \{\mathcal{W}(\mathbf{F}(\mathbf{x}), \mathbf{p}(\mathbf{x}), \mathbf{m}(\mathbf{x}), \mathbf{x}) - [\mathbf{e}_0(\mathbf{x}) \cdot \mathbf{p}(\mathbf{x}) + \mathbf{h}_0(\mathbf{x}) \cdot \mathbf{m}(\mathbf{x})]\}\, dv_x, \tag{2.71}$$

$$\begin{aligned} \mathcal{G}(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) &\equiv \tfrac{1}{2} \int_{\mathbf{R}^3} |\hat{E}_r(\mathbf{p}, \hat{\mathbf{y}}; \mathbf{y})|^2 + |\hat{H}_r(\mathbf{m}, \hat{\mathbf{y}}; \mathbf{y})|^2 dv_y \\ &= -\tfrac{1}{2} \int_\Omega [\hat{\mathbf{e}}_r(\mathbf{p}, \hat{\mathbf{y}}; \mathbf{x}) \cdot \mathbf{p}(\mathbf{x}) + \hat{\mathbf{h}}_r(\mathbf{m}, \hat{\mathbf{y}}; \mathbf{x}) \cdot \mathbf{m}(\mathbf{x})] dv_x. \end{aligned} \tag{2.72}$$

Accordingly we define

$$\mathcal{E}(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \equiv \tilde{\mathcal{E}}(\hat{\mathbf{y}}, \hat{E}_r(\mathbf{p}, \hat{\mathbf{y}}; \cdot), \mathbf{p}, \hat{H}_r(\mathbf{m}, \hat{\mathbf{y}}; \cdot), \mathbf{m}). \tag{2.73}$$

In getting the energy functional $\mathcal{E}$ we have made the trade-off indicated at the outset: The new energy is simpler because our unknowns are defined only on the reference configuration of the body; but it is also more complicated because the energy density now depends (through $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{h}}_r$) on the global values of the unknowns.

We now adapt this final form of the energy to the ferromagnetic materials considered in this paper. Adopting the constitutive assumptions (2.49) and (2.50), we consider an energy of the form

$$\begin{aligned} \check{\mathcal{E}}(\hat{\mathbf{y}}, \mathbf{m}) = \int_\Omega \{&\mathcal{V}(\mathbf{F}(\mathbf{x}), \mathbf{m}(\mathbf{x}), \mathbf{x}) - \mathbf{h}_0(\mathbf{x}) \cdot \mathbf{m}(\mathbf{x}) \\ &+ \mathcal{X}(\mathbf{F}(\mathbf{x}), \operatorname{Grad} \mathbf{m}(\mathbf{x}))\} dv_x + \mathcal{G}(\hat{\mathbf{y}}, \mathbf{m}), \end{aligned} \tag{2.74}$$

where the nonlocal energy $\mathcal{G}$ has been modified in the obvious way to account for our neglect of electric effects.

2.9. Minimization problems. Note that for either type of material, the nonlocal form of the energy, $\mathcal{E}$ and $\check{\mathcal{E}}$, respectively, is coercive. (This assumes, of course, that we put reasonable growth conditions on $\mathcal{W}$ and $\mathcal{V}$, as we do in §4.) Thus, it is reasonable to pose minimization problems. We begin with monotone materials.

Problem 2.4. Given $\mathbf{y}_0 \in W^{1,p}(\Omega)$ satisfying (2.3) and (2.4), $\mathbf{s}_0 \in L^{\frac{p}{p-1}}(S_2)$, and (E_0, H_0) satisfying (2.5); find $(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \in \mathcal{B}$ such that

$$I(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \equiv \mathcal{E}(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) - \int_{S_2} s_0 \cdot \hat{\mathbf{y}} da,$$

is minimized.

It follows from Brown [6] and Ciarlet and Nečas [8] that sufficiently regular solutions of the minimization Problem 2.4 combined with the solution operators $\hat{E}_r$ and $\hat{H}_r$ solve the variational Problem 2.3. Furthermore, it follows from the work of Brown [6], [5] that sufficiently regular solutions of the weak form of the Euler-Lagrange equations for the functional I,

$$\delta \mathcal{E}(\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) = \int_{S_2} s_0(\mathbf{x}) \cdot \delta \hat{\mathbf{y}}(x) da_x, \tag{2.75}$$

again combined with the electric and magnetic fields given by E_r and H_r, solve the local Problem 2.2. However, regularity theory for solutions to such problems remains incomplete. In particular, conditions such as (2.3) pose serious difficulties.

We now consider ferromagnetic materials. We begin by modifying the admissible states of the body to account for our omission of electric effects: Let

$$\tilde{\mathcal{B}} = \left\{ (\hat{\mathbf{y}}, \mathbf{m}) \;\middle|\; \begin{array}{l} \mathbf{m} \in W^{1,s}(\Omega), \quad s > \frac{2pr}{p(r+1)+2r}, \\ \text{such that there exists } \mathbf{p} \text{ with } (\hat{\mathbf{y}}, \mathbf{p}, \mathbf{m}) \in \mathcal{B} \end{array} \right\}. \tag{2.76}$$

Our problem can then be stated as follows.

Problem 2.5. Given $\mathbf{y}_0 \in W^{1,p}(\Omega)$ satisfying (2.3) and (2.4), $\mathbf{s}_0 \in L^{\frac{p}{p-1}}(\mathcal{S}_2)$, and H_0 satisfying (2.5), find $(\hat{\mathbf{y}}, \mathbf{m}) \in \tilde{\mathcal{B}}$ such that

$$\check{I}(\hat{\mathbf{y}}, \mathbf{m}) \equiv \check{\mathcal{E}}(\hat{\mathbf{y}}, \mathbf{m}) - \int_{\mathcal{S}_2} \mathbf{s}_0 \cdot \hat{\mathbf{y}} da \tag{2.77}$$

is minimized.

3. Continuity and compactness results. In this section we give some results on the convergence of the solution operators defined above for various sequences of data.

It follows directly from (2.54) that the spatial version of the solution operator is weakly continuous.

THEOREM 3.1. *For any sequence of polarizations* $\{P^j\}$ *and magnetizations* $\{M^j\}$ *such that*

$$\left.\begin{array}{l} P^j \rightharpoonup \bar{P} \\ M^j \rightharpoonup \bar{M} \end{array}\right\} \text{ in } L^2(\mathbf{R}^3),$$

it follows that

$$\left.\begin{array}{l} \check{E}(P^j; \cdot) \rightharpoonup \check{E}(\bar{P}, \cdot) \\ \check{H}(M^j; \cdot) \rightharpoonup \check{H}(\bar{M}, \cdot) \end{array}\right\} \text{ in } L^2(\mathbf{R}^3),$$

where the half arrow $\rightharpoonup$ *indicates weak convergence.*

The situation in material coordinates is not so simple. Here, we must consider composite limits of deformations, polarizations, and magnetizations. The following lemma is useful in this context.

LEMMA 3.2. *For any sequence* $\{\hat{\mathbf{y}}^j, G^j\}$ *such that*

$$\det \mathbf{F}^j > 0 \text{ a.e.}, \tag{3.1}$$

$$\int_\Omega \det \mathbf{F}^j \leq |\hat{\mathbf{y}}^j(\Omega)|, \tag{3.2}$$

$$\text{spt } G^j \subset \hat{\mathbf{y}}^j(\Omega), \tag{3.3}$$

and

$$\hat{\mathbf{y}}^j \rightharpoonup \bar{\mathbf{y}} \quad \text{in } W^{1,p}(\Omega), \tag{3.4}$$

$$\det \mathbf{F}^j \rightharpoonup \det \bar{\mathbf{F}} \quad \text{in } L^r(\Omega), \tag{3.5}$$

$$G^j \rightharpoonup \bar{G} \quad \text{in } L^q(\mathbf{R}^3) \tag{3.6}$$

for some $p \geq 2$, $r > 1$, $q > 1$ *it follows that*

$$\det \mathbf{F}^j(\cdot) G^j(\hat{\mathbf{y}}^j(\cdot)) \rightharpoonup \det \bar{\mathbf{F}}(\cdot) \bar{G}(\bar{\mathbf{y}}(\cdot)) \text{ in } L^{q'}(\Omega) \tag{3.7}$$

with $q' = \frac{qr}{r+q-1}$.

Proof. Using Hölder's inequality, (3.5), (3.6), and the boundedness of weakly convergent sequences, we get

$$\begin{aligned}
&\int_\Omega |\det \mathbf{F}^j(\mathbf{x}) G^j(\hat{\mathbf{y}}^j(\mathbf{x}))|^{q'} dv_x \\
&\leq \left\{\int_\Omega |\det \mathbf{F}^j(\mathbf{x})|^r dv_x\right\}^{\frac{q-q'}{q}} \left\{\int_\Omega |G^j(\hat{\mathbf{y}}^j(\mathbf{x}))|^q \det \mathbf{F}^j(\mathbf{x}) dv_x\right\}^{\frac{q'}{q}} \\
&\leq \left\{\int_\Omega |\det \mathbf{F}^j(\mathbf{x})|^r dv_x\right\}^{\frac{q-q'}{q}} \left\{\int_{\mathbf{R}^3} |G(\mathbf{y})|^q \det dv_y\right\}^{\frac{q'}{q}} \\
&\leq C.
\end{aligned} \tag{3.8}$$

Thus, by the weak compactness of bounded sets in $L^{q'}$, there exists an $f \in L^{q'}(\Omega)$ such that at least for a subsequence

$$\det \mathbf{F}^j(\cdot) G^j(\hat{\mathbf{y}}^j(\cdot)) \rightharpoonup f(\cdot) \quad \text{in } L^{q'}(\Omega).$$

We now show that

$$f(\cdot) = \det \bar{\mathbf{F}}(\cdot) \bar{G}(\bar{\mathbf{y}}(\cdot)).$$

Let $\hat{\mathbf{x}}^j$ and $\bar{\mathbf{x}}$ denote the respective inverses of $\hat{\mathbf{y}}^j$ and $\bar{\mathbf{y}}$. (The invertibility of $\hat{\mathbf{y}}^j$ follows from (3.1) and (3.2); the invertibility of $\bar{\mathbf{y}}$ is shown in [8].) Now for any $\phi \in C(\Omega)$ we define the spatial fields

$$\phi^j(\mathbf{y}) = \begin{cases} \phi(\hat{\mathbf{x}}^j(\mathbf{y})), & \mathbf{y} \in \hat{\mathbf{y}}^j(\Omega), \\ 0 & \text{elsewhere;} \end{cases}$$

and

$$\bar{\phi}(\mathbf{y}) = \begin{cases} \phi(\bar{\mathbf{x}}(\mathbf{y})), & \mathbf{y} \in \bar{\mathbf{y}}(\Omega), \\ 0 & \text{elsewhere.} \end{cases}$$

Since $\hat{\mathbf{x}}^j \to \bar{\mathbf{x}}$ almost everywhere (at least for a subsequence) and ϕ is continuous we have

$$\phi^j \to \bar{\phi} \quad \text{a.e. in } \mathbf{R}^3.$$

Also, for any $s \in [1, \infty)$ we have

$$\begin{aligned}
\int_{\mathbf{R}^3} |\phi^j(\mathbf{y})|^s dv_y &= \int_\Omega |\phi(\mathbf{x})|^s \det \mathbf{F}^j(\mathbf{x}) dv_x \\
&\to \int_\Omega |\phi(\mathbf{x})|^s \det \bar{\mathbf{F}}(\mathbf{x}) dv_x \\
&= \int_{\mathbf{R}^3} |\bar{\phi}(\mathbf{y})|^s dv_y.
\end{aligned}$$

It follows from a standard exercise in real variable theory that

$$\phi^j \to \bar{\phi} \quad \text{(strongly) in } L^s(\mathbf{R}^3).$$

Thus, for any $\phi \in C(\Omega)$

$$\begin{aligned}
\int_\Omega \det \mathbf{F}^j(\mathbf{x}) G^j(\hat{\mathbf{y}}^j(\mathbf{x})) \phi(\mathbf{x}) dv_x &= \int_{\mathbf{R}^3} G^j(\mathbf{y}) \phi^j(\mathbf{y}) dv_y \\
&\to \int_{\mathbf{R}^3} \bar{G}(\mathbf{y}) \bar{\phi}(\mathbf{y}) dv_y \\
&= \int_\Omega \det \bar{\mathbf{F}}(\mathbf{x}) \bar{G}(\bar{\mathbf{y}}(\mathbf{x})) \phi(\mathbf{x}) dv_x.
\end{aligned}$$

Here we have used the fact that the product of a weakly convergent and a strongly convergent sequence (in the appropriate spaces) converges weakly in the sense of distributions. The conclusion of the lemma follows directly from a density argument and the uniqueness of weak limits. □

The following theorem on the limits of material functions will be used in our existence theorem for monotone magnetic materials.

THEOREM 3.3. *Let* $\{\hat{\mathbf{y}}^j, \mathbf{p}^j, \mathbf{m}^j\} \subset \mathcal{B}$ *satisfy*

$$\|\mathbf{F}^j\|_{L^p(\Omega)} + \|\mathbf{F}^{j\times}\|_{L^q(\Omega)} + \|\det \mathbf{F}^j\|_{L^r(\Omega)} + \left\| \frac{\mathbf{F}^j \cdot \mathbf{p}^j}{\sqrt{\det \mathbf{F}^j}} \right\|_{L^2(\Omega)} + \left\| \frac{\mathbf{F}^j \cdot \mathbf{m}^j}{\sqrt{\det \mathbf{F}^j}} \right\|_{L^2(\Omega)} \leq C, \tag{3.9}$$

and

$$\int_\Omega \mathcal{W}(\mathbf{F}^j(\mathbf{x}), \mathbf{p}^j(\mathbf{x}), \mathbf{m}^j(\mathbf{x}), x) dv_x \leq C; \tag{3.10}$$

then there exists $(\bar{\mathbf{y}}, \bar{\mathbf{p}}, \bar{\mathbf{m}}) \in \mathcal{B}$ *such that (at least for a subsequence)*

$$\hat{\mathbf{y}}^j \rightharpoonup \bar{\mathbf{y}} \quad \text{in } W^{1,p}(\Omega), \tag{3.11}$$

$$\mathbf{F}^{j\times} \rightharpoonup \bar{\mathbf{F}}^\times \quad \text{in } L^q(\Omega), \tag{3.12}$$

$$\det \mathbf{F}^j \rightharpoonup \det \bar{\mathbf{F}} \quad \text{in } L^r(\Omega), \tag{3.13}$$

$$\mathbf{F}^j \cdot \mathbf{p}^j \rightharpoonup \bar{\mathbf{F}} \cdot \bar{\mathbf{p}} \quad \text{in } L^s(\Omega), \tag{3.14}$$

$$\mathbf{F}^j \cdot \mathbf{m}^j \rightharpoonup \bar{\mathbf{F}} \cdot \bar{\mathbf{m}} \quad \text{in } L^s(\Omega), \tag{3.15}$$

$$\hat{E}(\hat{\mathbf{y}}^j, \mathbf{p}^j; \cdot) \rightharpoonup \hat{E}(\bar{\mathbf{y}}, \bar{\mathbf{p}}; \cdot) \quad \text{in } L^2(\mathbf{R}^3), \tag{3.16}$$

$$\hat{H}(\hat{\mathbf{y}}^j, \mathbf{m}^j; \cdot) \rightharpoonup \hat{H}(\bar{\mathbf{y}}, \bar{\mathbf{m}}; \cdot) \quad \text{in } L^2(\mathbf{R}^3), \tag{3.17}$$

where $s = \frac{2r}{r+1}$.

Proof. The existence of a $\bar{\mathbf{y}}$, suitable as a component of an element of $\mathcal{B}$, such that (3.11), (3.12), and (3.13) hold, follows from the work of Ball [3] and the refinements of Ciarlet and Nečas [8]. To deduce the existence of $\bar{\mathbf{p}}$ and $\bar{\mathbf{m}}$ we define

$$P^j(\mathbf{y}) = \tilde{P}(\hat{\mathbf{x}}^j, \mathbf{p}^j; \mathbf{y}) \tag{3.18}$$

and

$$M^j(\mathbf{y}) = \tilde{M}(\hat{\mathbf{x}}^j, \mathbf{m}^j; \mathbf{y}) \tag{3.19}$$

where $\tilde{P}$ and $\tilde{M}$ are defined in (2.60) and (2.61), respectively. We note that

$$\|P^j\|_{L^2(\mathbf{R}^3)} = \left\| \frac{\mathbf{F}^j \cdot \mathbf{p}^j}{\sqrt{\det \mathbf{F}^j}} \right\|_{L^2(\Omega)}, \quad \text{and}$$

$$\|M^j\|_{L^2(\mathbf{R}^3)} = \left\| \frac{\mathbf{F}^j \cdot \mathbf{m}^j}{\sqrt{\det \mathbf{F}^j}} \right\|_{L^2(\Omega)}.$$

Thus, by (3.9) and the weak compactness of bounded sets in L^2, there exists $(\bar{P}, \bar{M})$ such that (at least for a subsequence)

$$P^j \rightharpoonup \bar{P} \quad \text{in } L^2(\mathbf{R}^3) \quad \text{and } M^j \rightharpoonup \bar{M} \quad \text{in } L^2(\mathbf{R}^3). \tag{3.20}$$

It follows from Theorem 3.1 that

$$\check{E}_r(P^j;\cdot) \rightharpoonup \check{E}_r(\bar{P};\cdot) \quad \text{and} \tag{3.21}$$

$$\check{H}_r(M^j;\cdot) \rightharpoonup \check{H}_r(\bar{M};\cdot). \tag{3.22}$$

If we define

$$\begin{aligned} \bar{\mathbf{p}} &= (\det \bar{\mathbf{F}})\bar{\mathbf{F}}^{-1} \cdot \bar{P}(\bar{\mathbf{y}}(\cdot)) \text{ and} \\ \bar{\mathbf{m}} &= (\det \bar{\mathbf{F}})\bar{\mathbf{F}}^{-1} \cdot \bar{M}(\bar{\mathbf{y}}(\cdot)), \end{aligned} \tag{3.23}$$

then (3.14)–(3.17) follow directly from (2.67), (2.68), (3.20), and Lemma 3.2. Thus, our theorem is proved. □

The final result of this section is useful in our theorems on classical ferromagnetic materials.

THEOREM 3.4. *If for some sequence* $\{\hat{\mathbf{y}}^j, \mathbf{m}^j\}$ *such that*

$$\det \mathbf{F}^j > 0 \text{ a.e.}, \tag{3.24}$$

$$\int_\Omega \det \mathbf{F}^j \leq |\hat{\mathbf{y}}^j(\Omega)|, \tag{3.25}$$

$$\left\| \frac{\mathbf{F}^j \cdot \mathbf{m}^j}{\sqrt{\det \mathbf{F}^j}} \right\|_{L^2(\Omega)} \leq C, \tag{3.26}$$

and for some $r > 1$, $p > \frac{2r}{r-1}$, $s > \frac{2pr}{p(r+1)+2r}$

$$\hat{\mathbf{y}}^j \rightharpoonup \bar{\mathbf{y}} \quad \textit{(weakly) in } W^{1,p}(\Omega), \tag{3.27}$$

$$\det \mathbf{F}^j \rightharpoonup \det \bar{\mathbf{F}} \quad \textit{(weakly) in } L^r(\Omega), \textit{ and} \tag{3.28}$$

$$\mathbf{m}^j \to \bar{\mathbf{m}} \quad \textit{(strongly) in } L^s(\Omega), \tag{3.29}$$

then $\bar{\mathbf{y}}$ *is invertible with inverse* $\bar{\mathbf{x}}$, *and if* M^j *is defined as in* (3.19) *and*

$$\bar{M}(\mathbf{y}) \equiv \begin{cases} [\det \bar{\mathbf{F}}(\bar{\mathbf{x}}(\mathbf{y}))]^{-1}\bar{\mathbf{F}}(\bar{\mathbf{x}}(\mathbf{y})) \cdot \bar{\mathbf{m}}(\bar{\mathbf{x}}(\mathbf{y})), & \mathbf{y} \in \bar{\mathbf{y}}(\Omega), \\ \mathbf{0} & \textit{elsewhere} \end{cases} \tag{3.30}$$

then it follows that

$$M^j \rightharpoonup \bar{M} \quad \textit{in } L^2(\mathbf{R}^3), \tag{3.31}$$

and

$$\hat{H}_r(\mathbf{y}^j, \mathbf{m}^j, \cdot) \rightharpoonup \hat{H}_r(\bar{\mathbf{y}}, \bar{\mathbf{m}}, \cdot) \quad \textit{in} \quad L^2(\mathbf{R}^3). \tag{3.32}$$

Proof. As in the previous theorem, (3.26) implies that there exists $\check{M} \in L^2(\mathbf{R}^3)$ such that at least for a subsequence

$$M^j \rightharpoonup \check{M} \text{ in } L^2(\mathbf{R}^3).$$

We must show that $\bar{M} = \check{M}$. We first observe that Lemma 3.2 implies that at least for a subsequence

$$\mathbf{F}^j \cdot \mathbf{m}^j = (\det \mathbf{F}^j) M^j(\mathbf{y}^j(\cdot)) \rightharpoonup (\det \bar{\mathbf{F}}) \check{M}(\bar{\mathbf{y}}(\cdot)) \quad \text{in } L^{q'}(\Omega)$$

with $q' = \frac{2r}{r+1}$. But (3.27) and (3.29) imply

$$\mathbf{F}^j \cdot \mathbf{m}^j \rightharpoonup \bar{\mathbf{F}} \cdot \bar{\mathbf{m}} \quad \text{in } L^q(\Omega) \tag{3.33}$$

with $q = \frac{ps}{p+s} > q'$. Thus, by (3.30) and the uniqueness of weak limits we have $\bar{M} = \check{M}$ and (3.31) is proved. It follows from Theorem 3.1 that

$$\check{H}_r(M^j; \cdot) \rightharpoonup \check{H}_r(\bar{M}; \cdot), \tag{3.34}$$

and (3.32) follows from (2.68). □

4. Existence. We state and prove an existence theorem for the minimization Problem 2.4 for monotone materials.

THEOREM 4.1. *Let* $\mathbf{y}_0 \in W^{1,p}(\Omega)$ *satisfying* (2.3) *and* (2.4), $\mathbf{s}_0 \in L^{\frac{p}{p-1}}(S_2)$, *and* (E_0, H_0) *satisfying* (2.5) *be given. Let* $\mathcal{W}$ *be* polyconvex, *i.e., let it satisfy*

$$\mathcal{W}(\mathbf{F}, \mathbf{p}, \mathbf{m}, \mathbf{x}) = \Upsilon(\mathbf{F}, \mathbf{F}^\times, \det \mathbf{F}, \mathbf{F} \cdot \mathbf{p}, \mathbf{F} \cdot \mathbf{m}, \mathbf{x}), \tag{4.1}$$

with $\Upsilon(\cdot, \cdot, \cdot, \cdot, \cdot, \mathbf{x})$ *convex for each* $\mathbf{x} \in \Omega$. *(See Ball* [3] *for the definition of polyconvexity in a purely elastic material.) Let* Υ *satisfy*

$$\Upsilon(\mathbf{F}, \mathbf{F}^\times, \delta, \mathbf{F} \cdot \mathbf{p}, \mathbf{F} \cdot \mathbf{m}, \mathbf{x}) \to \infty \text{ as } \delta \to 0, \tag{4.2}$$

and suppose that there exist numbers $k > 0$, $p \geq 2$, $q > p/(p-1)$, $r > 1$, *and a function* $\omega \in L^1(\Omega)$ *such that*

$$\Upsilon(\mathbf{F}, \mathbf{F}^\times, \delta, \mathbf{F}\cdot\mathbf{p}, \mathbf{F}\cdot\mathbf{m}, \mathbf{x}) \geq \omega(\mathbf{x}) + k \left(|\mathbf{F}|^p + |\mathbf{F}^\times|^q + \delta^r + \frac{|\mathbf{F} \cdot \mathbf{p}|^2}{\det \mathbf{F}} + \frac{|\mathbf{F} \cdot \mathbf{m}|^2}{\det \mathbf{F}} \right) \tag{4.3}$$

for every $\mathbf{x} \in \Omega$. *Suppose further that there exists an element* $(\hat{\mathbf{y}}_1, \mathbf{p}_1, \mathbf{m}_1) \in \mathcal{B}$ *such that* $I(\hat{\mathbf{y}}_1, \mathbf{p}_1, \mathbf{m}_1) < \infty$. *Then there exists a solution* $(\bar{\mathbf{y}}, \bar{\mathbf{p}}, \bar{\mathbf{m}}) \in \mathcal{B}$ *to Problem* 2.4.

Proof. Since I is bounded below there exists an infimizing sequence $(\hat{\mathbf{y}}^j, \mathbf{p}^j, \mathbf{m}^j) \subset \mathcal{B}$. It follows from Theorem 3.3 that there exists $(\bar{\mathbf{y}}, \bar{\mathbf{p}}, \bar{\mathbf{m}}) \in \mathcal{B}$ such that (3.11)–(3.17) hold. The weak lower-semicontinuity of $\mathcal{L}$ follows from the convexity of Υ and Tonelli's theorem (cf. [9, p.7]). The weak lower-semicontinuity of $\mathcal{G}$ follows from (3.16), (3.17), and Tonelli's theorem. Thus

$$I(\bar{\mathbf{y}}, \bar{\mathbf{p}}, \bar{\mathbf{m}}) \leq \liminf I(\hat{\mathbf{y}}^j, \mathbf{p}^j, \mathbf{m}^j). \tag{4.4}$$

So $I(\bar{\mathbf{y}}, \bar{\mathbf{p}}, \bar{\mathbf{m}})$ is minimal, and our theorem is proved. □

We now prove an existence theorem for the minimization Problem 2.5 for ferromagnetic materials. Our result is a generalization of that of Visintin [18] which considered ferromagnetism of rigid bodies. The proof is very similar to that of Theorem 4.1, but here we use the additional compactness in $\mathbf{m}$ given by the exchange energy $\mathcal{X}$ to compensate for the nonconvexity of the stored energy $\mathcal{V}$. To do this we employ the following lemma.

LEMMA 4.2. *We consider the function*

$$\mathbf{R}^m \times \mathbf{R}^l \times \Omega \ni (u, v, \mathbf{x}) \mapsto F(u, v, \mathbf{x}) \in \mathbf{R}. \tag{4.5}$$

We assume that $\mathbf{F}$ *is measurable in* $\mathbf{x}$, *continuous in its other arguments, and convex in* u. *Then if* (u^j, v^j) *is a sequence of functions on* Ω *such that*

$$u^j \rightharpoonup \bar{u} \qquad \text{in } L^1(\Omega), \tag{4.6}$$

$$v^j \to \bar{v} \qquad \text{a.e in } (\Omega), \tag{4.7}$$

and

$$F(u^j(\mathbf{x}), v^j(\mathbf{x}), \mathbf{x}) \geq f(\mathbf{x}), \quad F(\bar{u}(\mathbf{x}), \bar{v}(\mathbf{x}), \mathbf{x}) \geq f(\mathbf{x}) \tag{4.8}$$

for some $f \in L^1(\Omega)$, *then*

$$\int_\Omega F(\bar{u}(\mathbf{x}), \bar{v}(\mathbf{x}), \mathbf{x}) \leq \liminf_{j\to\infty} \int_\Omega F(u^j(\mathbf{x}), v^j(\mathbf{x}), \mathbf{x}). \tag{4.9}$$

The proof of this can be found in Eisen [10] and Ball, Currie, and Olver [4].

THEOREM 4.3. *Let* $\mathbf{y}_0 \in W^{1,p}(\Omega)$ *satisfying* (2.3) *and* (2.4), $\mathbf{s}_0 \in L^{\frac{p}{p-1}}(S_2)$, *and* H_0 *satisfying* (2.5) *be given. Let* $\mathcal{V}$ *be measurable in* $\mathbf{x}$, *continuously differentiable in its other arguments, and satisfy*

$$\mathcal{V}(\mathbf{F}, \mathbf{m}, \mathbf{x}) = \tilde{\Upsilon}(\mathbf{F}, \mathbf{F}^\times, \det \mathbf{F}, \mathbf{m}, \mathbf{x}), \tag{4.10}$$

with $\Upsilon(\cdot, \cdot, \cdot, \mathbf{m}, \mathbf{x})$ *convex for each* $(\mathbf{m}, \mathbf{x}) \in \mathbf{R}^3 \times \Omega$. *Assume also that*

$$\tilde{\Upsilon}(\mathbf{F}, \mathbf{F}^\times, \delta, \mathbf{m}, \mathbf{x}) \to \infty \text{ as } \delta \to 0; \tag{4.11}$$

that there exist numbers $k_1 > 0$, $r > 1$, $p \geq \frac{2r}{r-1}$, $q > p/(p-1)$, $s > \frac{2pr}{p(r+1)+2r}$, *and a function* $\omega \in L^1(\Omega)$ *such that*

$$\tilde{\Upsilon}(\mathbf{F}, \mathbf{F}^\times, \delta, \mathbf{m}, \mathbf{x}) \geq \omega(\mathbf{x}) + k_1 \left(|\mathbf{F}|^p + |\mathbf{F}^\times|^q + \delta^r + |\mathbf{m}|^s + \frac{|\mathbf{F}\cdot\mathbf{m}|^2}{\det \mathbf{F}} \right) \tag{4.12}$$

for every $\mathbf{x} \in \Omega$. *Assume further that the exchange energy density satisfies*

$$\mathcal{X}(\mathbf{F}, \mathbf{G}) = \tilde{\mathcal{X}}(\mathbf{F}, \mathbf{F}^\times, \det \mathbf{F}, \mathbf{G}), \tag{4.13}$$

where $\tilde{\mathcal{X}}(\cdot, \cdot, \cdot, \cdot)$ *is convex and*

$$\mathcal{X}(\mathbf{F}, \mathbf{G}) \geq k_2 + k_3(|\mathbf{F}|^p + |\mathbf{F}^\times|^q + |\det \mathbf{F}|^r + |\mathbf{G}|^s), \tag{4.14}$$

for some positive constants k_2 *and* k_3. *Then if there exists an element* $(\hat{\mathbf{y}}_1, \mathbf{m}_1) \in \tilde{\mathcal{B}}$ *such that* $\check{I}(\hat{\mathbf{y}}_1, \mathbf{m}_1) < \infty$, *there exists a solution* $(\bar{\mathbf{y}}, \bar{\mathbf{m}}) \in \tilde{\mathcal{B}}$ *to Problem* 2.5.

Proof. As usual, let $\{\hat{\mathbf{y}}^j, \mathbf{m}^j\}$ be an infimizing sequence for $\check{I}$. The growth conditions (4.12) and (4.14) imply that

$$\|\mathbf{m}^j\|_{W^{1,s}(\Omega)} < C.$$

Thus, there exists $\bar{\mathbf{m}} \in W^{1,s}(\Omega)$ such that (at least for a subsequence)

$$\mathbf{m}^j \rightharpoonup \bar{\mathbf{m}} \quad \text{in } W^{1,s}(\Omega), \tag{4.15}$$

and

$$\mathbf{m}^j \to \bar{\mathbf{m}} \quad \text{in } L^s(\Omega). \tag{4.16}$$

As before, there also exists $\bar{\mathbf{y}} \in W^{1,p}(\Omega)$ such that

$$\begin{aligned} &\hat{\mathbf{y}}^j \rightharpoonup \bar{\mathbf{y}} \quad \text{in } W^{1,p}(\Omega), \\ &\mathbf{F}^{j\times} \rightharpoonup \bar{\mathbf{F}}^{\times} \quad \text{in } L^q(\Omega), \text{ and} \\ &\det \mathbf{F}^j \rightharpoonup \det \bar{\mathbf{F}} \quad \text{in } L^r(\Omega). \end{aligned}$$

Thus, it follows from (3.32) of Theorem 3.4, the convexity of $\mathcal{G}$ (in its spatial form), and Tonelli's theorem that

$$\mathcal{G}(\bar{\mathbf{F}}, \bar{\mathbf{m}}) \leq \liminf \mathcal{G}(\mathbf{F}^j, \mathbf{m}^j).$$

The weak lower-semicontinuity of $\mathcal{V}$ follows from Lemma 4.2, and the weak lower-semicontinuity of the other terms follows as in the proof of Theorem 4.1. □

As a simple example of an energy density satisfying our hypotheses, consider

$$\begin{aligned} \mathcal{V}(\mathbf{F}, \mathbf{m}, \mathbf{x}) \quad &= \quad \mathcal{W}(\mathbf{F}, \mathbf{F}^{\times}, \det \mathbf{F}, \mathbf{x}) + \kappa_1 \left(\frac{|\mathbf{m}|^4}{4} - m_0^2 \frac{|\mathbf{m}|^2}{2} + \frac{m_0^4}{4} \right)^{\frac{s}{4}} \\ &\quad + \kappa_2 \frac{|\mathbf{F} \cdot \mathbf{m}|^2}{(\det \mathbf{F})}. \end{aligned} \tag{4.17}$$

Here κ_1 and κ_2 are positive constants, and we take $\mathcal{W}$ to be polyconvex. Thus the first term represents a purely elastic part of the energy. The second term describes the pure magnetic part (this would be the energy of a rigid ferromagnetic material with preferred magnetization m_0); and the third term ensures that the energy of the spatial magnetic field is bounded above. We could of course introduce other coupled terms modeling such effects as magnetostriction.

5. Comments. We conclude with some remarks on possible extensions of these results.

1. In the linear theory of *diamagnetic materials* one assumes a constitutive relation of the form

$$M = \chi H$$

with χ negative but $|\chi| < 1$. Thus, while map $H \mapsto M$ is "monotone decreasing," the map $H \mapsto \mathfrak{b}$ where

$$\mathfrak{b} \equiv H + M = H + \chi H$$

is monotone (increasing).

The results of [15], where we chose $\mathbf{h}$ as an independent variable and $\mathbf{b}$ as a dependent variable, depended on the monotonicity of $\mathbf{b}$ without regard to $\mathbf{m}$, and thus applied to diamagnetic materials. In this regard, we note that one should be able to obtain an existence result for materials for which the map $\mathbf{m} \mapsto \mathcal{W}$ is *concave* from the results above as long as proper coercivity assumptions are placed on the entire energy.

2. In ferroelectric materials one usually assumes that the stored energy is nonconvex, but that unlike ferromagnetic materials there is no exchange energy. Thus, we can expect neither the weak continuity found in the problems for monotone materials nor the compactness given by the exchange energy in the problems for ferromagnetic materials. It may instead be possible to consider "measure valued" solutions such as

those considered in the theory of twinning in crystals as developed by Ericksen [11], and Chipot and Kinderlehrer [7]. The idea here is that while the weak limit of an infimizing sequence may not minimize a nonconvex energy functional, the Young's measure for the sequence (a probability measure whose center of mass is the weak limit) may represent the physical situation that minimizes energy. It may be possible to model the microscopic structure of the domains (highly discontinuous regions of uniform polarization) found in ferromagnetic materials by a macroscopic Young's measure indicating the distribution of the domains.

Acknowledgments. The author would like to thank the referees of this paper and John Ball for their helpful suggestions.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] J. F. G. AUCHMUTY AND R. BEALS, *Variational solutions of some nonlinear free boundary problems*, Arch. Rat. Mech. Anal., 43(1971), pp. 255–271.
[3] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rat. Mech. Anal., 63(1977), pp. 337–403.
[4] J. M. BALL, J. C. CURRIE, AND P. J. OLVER, *Null Lagrangians, weak continuity, and variational problems of arbitrary order*, J. Funct. Anal., 41(1981), pp. 135–174.
[5] W. F. BROWN, *Magnetoelastic Interactions*, Springer, Berlin, 1966.
[6] ——, *Magnetostatic Principles in Ferromagnetism*, North-Holland, Amsterdam, 1962.
[7] M. CHIPOT AND D. KINDERLEHRER, *Equilibrium Configurations of Crystals*, Technical Report 326, Institute for Mathematics and its Applications, 1987.
[8] P. G. CIARLET AND J. NEČAS, *Injectivity and self-contact in nonlinear elasticity*, Arch. Rat. Mech. Anal., 97(1987), pp. 171–188.
[9] B. DACOROGNA, *Weak continuity and weak lower semi-continuity of nonlinear functionals*, Lecture Notes in Mathematics 922, Springer-Verlag, Berlin, 1982.
[10] G. EISEN, *A selection lemma for sequences of measurable sets, and lower semicontinuity of multiple integrals*, Manuscripta Math., 27(1979), pp. 73–79.
[11] J. L. ERICKSEN, *Twinning of crystals* I, in Metastability and Incompletely Posed Problems, S. S. Antman, J.L. Ericksen, D. Kinderlehrer, and I. Müller, eds., IMA, Springer-Verlag, New York, 1986.
[12] K. HUTTER AND A. A. F. VAN DE VEN, *Field-matter interactions in thermoelastic solids*, Lecture Notes in Physics 88, Springer-Verlag, Berlin, 1978.
[13] O. D. KELLOGG, *Foundations of Potential Theory*, Springer, Berlin, 1929.
[14] L. D. LANDAU AND E. M. LIFSHITZ, *Electrodynamics of Continuous Media*, Pergamon Press, 1960.
[15] R. C. ROGERS AND S. S. ANTMAN, *Steady-state problems of nonlinear electro-magneto-thermo-elasticity*, Arch. Rat. Mech. Anal., 95(1986), pp. 279–323.
[16] H. F. TIERSTEN, *Variational principle for saturated magnetoelastic insulators*, J. Math. Phys., 6(1965), pp. 779–787.
[17] R. A. TOUPIN, *The elastic dielectric*, Arch. Rat. Mech. Anal., 5(1956), pp. 849–915.
[18] A. VISINTIN, *On Landau-Lifshitz' equations for ferromagnetism*, Japan J. Appl. Math., 2(1985), pp. 69–84.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

007

THE INTEGRAL REPRESENTATION OF THE POSITIVE SOLUTIONS OF THE GENERALIZED WEINSTEIN EQUATION ON A QUARTER-SPACE*

ÖMER AKIN†

Abstract. The present paper gives a unique integral representation for the positive solutions to the generalized Weinstein equation

$$L[u]=L_{p,q}[u]\equiv\sum_{i=1}^{n}u_{x_ix_i}+\frac{p}{x_{n-1}}u_{x_{n-1}}+\frac{q}{x_n}u_{x_n}.$$

This integral representation is explicitly described in terms of hypergeometric functions. The methods used are those of potential theory; the technique of Martin plays a particularly crucial role.

Key words. potential theory, Martin boundary, Weinstein equation, integral representation, minimal harmonic functions

AMS(MOS) subject classifications. 31, 31B, 31D

1. Introduction.

The equation

$$L[u]=L_{p,q}[u]\equiv\sum_{i=1}^{n}u_{x_ix_i}+\frac{p}{x_{n-1}}u_{x_{n-1}}+\frac{q}{x_n}u_{x_n} \tag{1.1}$$

has been treated by Gilbert [1], [2] by integral operator methods, and fundamental solutions to (1.1) were given by Weinacht [3]. Kapilevich [4] and Celebi [5] have given mean value theorems for a class of equations (1.1). Hall, Quinn, and Weinacht [6] have also given some mean value theorems for (1.1). Quinn and Weinacht have given existence and uniqueness theorems covering all parameter ranges of p and q [7].

In (1.1) if $p=0$ then the equation reduces to the form $L_q[u]=0$, which has been treated in [8]-[12]. The method used by Brelot in [12] is essentially different from those used in the other papers. He was the first mathematician to treat the equation $L_q[u]=0$ by using Martin's technique [13]. Brelot described all the positive solutions of that equation, in terms of an integral kernel given by Huber in [9]. We would like to use the potential theoretic methods of Martin; however, the problem is significantly more complicated than the problem treated by Brelot [12], because the symmetries of a quarter-space are more sparse than for a half-space.

The papers [7] and [11] introduced fundamental solutions for the operator $L_{p,q}$. It will be important to identify the minimal positive fundamental solutions (or, in other words, the classical Green functions). We do this when $p, q<1$ and use the correspondence principle to identify the minimal positive fundamental solutions if p or q is bigger than one. (The correspondence principle is as follows: A function u is a solution of $L_{2-p,2-q}[u]=0$ if and only if $v=x_{n-1}^{1-p}\cdot x_n^{1-q}\cdot u$ is a solution of $L_{p,q}[v]=0$.) After we have done this we follow Martin's method in considering the ratio of these Green functions. Standard results from potential theory tell us how to obtain an integral kernel describing all positive harmonic functions [13]-[17]. We will also identify the Martin boundary points corresponding to minimal positive harmonic functions and

* Received by the editors August 20, 1987; accepted for publication (in revised form) December 15, 1987. This research was done while the author was visiting the Department of Mathematics, University of Edinburgh, Edinburgh, Scotland.

† Department of Mathematics, Faculty of Science and Arts, University of Firat, Elazig, Turkey.

so obtain a unique representation for every positive harmonic function in the domain Q:

$$Q=\{x\colon (x_1, x_2, \cdots, x_n)\in \mathbb{R}^n,\ x_{n-1}>0,\ x_n>0\}.$$

Our formula is reasonably explicit.

2. Potentials for $L_{p,q}$. Let us suppose a function $\hat{E}_{p,q}(x,\xi)$ is given as follows:

$$\frac{-\Gamma(k)}{\pi^{n/2}\Gamma(p/2)\Gamma(q/2)}\int_0^{\pi}\int_0^{\pi}\sigma^{-2k}\sin^{p-1}\theta_1\sin^{q-1}\theta_2\,d\theta_1\,d\theta_2$$

where

$$\sigma\equiv\left[|x-\xi|^2+4x_{n-1}\xi_{n-1}\sin^2\left(\frac{\theta_1}{2}\right)+4x_n\xi_n\sin^2\left(\frac{\theta_2}{2}\right)\right]^{1/2},$$

$$2k=n+p+q-2,\qquad p,\ q\geqq 1.$$

We shall use $E_{p,q}(x,\xi)$ obtained from $\hat{E}_{p,q}$ as follows:

$$E_{p,q}(x,\xi)=-\xi_{n-1}^{p}\xi_n^{q}\hat{E}_{p,q}(x,\xi). \tag{2.1}$$

It is already known that (2.1) is a fundamental solution of $L_{p,q}$ with pole ξ [7]. We also know that

$$E^1_{p,q}(x,\xi)=x_{n-1}^{1-p}\cdot x_n^{1-q}\cdot E_{2-p,2-q}(x,\xi) \tag{2.2}$$

is a fundamental solution of (1.1) when $p<1$, $q<1$ [7].

We want to describe an integral kernel which will yield all positive harmonic functions on the Q. To do this we will apply the Martin's method as described in [13]. To apply the technique found there it is necessary that our fundamental solutions be potentials.

We need to prove that $E_{p,q}(x,\xi)$ is a potential on the quarter-space Q. To do this it is enough to show $E^1_{p,q}$ is a potential for $p,\ q<1$, because the correspondence principle conserves harmonicity[1] (superharmonicity, subharmonicity) and also preserves the order structure.

Applying the maximum principle on relatively compact open subsets of Q, we see that it will suffice to prove that

$$E^1_{p,q}(x,\xi)\to 0\quad \text{as } x\to\partial Q\quad \text{or } \infty$$

where ∂Q is the boundary of the quarter-space Q.

Let us suppose ξ is a fixed point in Q:

$$\lim_{x\to\partial Q}E^1_{p,q}(x,\xi)=\lim_{x\to\partial Q}x_n^{1-p}x_n^{1-q}\cdot\lim_{x\to\partial Q}E_{2-p,2-q}(x,\xi). \tag{2.3}$$

It is clear that in (2.3) the first limit is zero. We need to understand the second limit. We can assume that there is a δ such that $|x-\xi|>\delta$. It is easily shown that

$$\overline{\lim_{x\to\partial Q}}E_{2-p,2-q}(x,\xi)<\pi^2\delta^{p+q-n-2}.$$

This result shows that in (2.3) the second limit is bounded from above. It follows that

$$\lim_{x\to\partial Q}E^1_{p,q}(x,\xi)=0.$$

[1] In this paper as in Brelot-Collin and Brelot's article [12] harmonic, superharmonic, etc. are used relative to L rather than Δ.

Now, we need to show

$$\lim_{|x|\to\infty} E_{p,q}(x,\xi)=0.$$

To show this, we use the fact that $\delta<|x-\xi|$ so

$$\int_0^{\pi}\int_0^{\pi}\sigma^{p+q-n-2}\sin^{1-p}\theta_1\sin^{1-q}\theta_2\,d\theta_1\,d\theta_2\leqq|x-\xi|^{p+q-n-2}\pi^2.$$

This estimate only holds because $p, q<1$. So,

$$\lim_{|x|\to\infty} E^1_{p,q}(x,\xi)\leqq\lim_{|x|\to\infty} x_{n-1}^{1-p}\cdot x_n^{1-q}|x-\xi|^{p+q-n-2}\cdot\pi^2.$$

We note that the convergence of the right-hand side is uniform in the direction of approaching ∞.

Thus, $E^1_{p,q}(x,\xi)$ is a potential in the quarter-space Q for $L_{p,q}$ if $p, q<1$. Consequently, $E(x,\xi)$ is a potential in the same space for $L_{p,q}$ if $p, q\geqq 1$. The function $E(x,\xi)$ is a Green function and Q, with E, is a Green space in the sense of potential theory [15].

3. The Martin boundary.

(a) The case where $p, q>1$. In the previous part we proved that $E_{p,q}$, defined at (2.1), is a Green function. In this part, we will construct the minimal harmonic functions for $L_{p,q}$ by considering limits of ratios of Green functions. In other words, we will use Martin's technique. To apply that technique we will prove that the ratio of Green functions $E(x,\xi)/E(x_0,\xi)$ has a continuous extension to the compact space $Q\cup\partial Q\cup\{\infty\}$ as a function of ξ for any x, x_0 in Q. We keep x_0 fixed and denote the extension of $E(x,\xi)/E(x_0,\xi)$ by $K(x,\Xi)$ where $\Xi\in\partial Q\cup\{\infty\}$ and $x\in Q$.

Let us find $K(x,\Xi)$

$$\lim_{\xi\to\Xi}\frac{E(x,\xi)}{E(x_0,\xi)}=\lim_{\xi\to\Xi}\frac{\hat{E}(x,\xi)}{\hat{E}(x_0,\xi)}=K(x,\Xi). \tag{3.1}$$

To calculate the limit (3.1) we will treat two cases: $\Xi\in\partial Q$ and $\Xi=\{\infty\}$. In the first case, without loss of generality, we may assume that Ξ_n, the nth coordinate of Ξ, is zero in (3.1); otherwise, we merely interchange the last two coordinates, namely,

$$\Xi=(\Xi_1,\Xi_2,\cdots,\Xi_{n-1},0).$$

It is easily seen that

$$\hat{E}(x,\xi)=2^{p+q-2}\frac{\Gamma(k)}{\pi^{n/2}}\cdot\frac{\Gamma(p/2)}{\Gamma(p)}\cdot\frac{\Gamma(q/2)}{\Gamma(q)}\cdot|x-\xi|^{-2k}\cdot F_A\left(k,\frac{p}{2},\frac{q}{2},p,q,-\varepsilon_1^{-2},-\varepsilon_2^{-2}\right)$$

where

$$(\varepsilon_1,\varepsilon_2)=\left(\frac{|x-\xi|}{\sqrt{4x_{n-1}\xi_{n-1}}},\frac{|x-\xi|}{\sqrt{4x_n\xi_n}}\right)$$

and F_A is the function of Lauricella [16]. So,

$$\lim_{\xi\to\Xi}\hat{E}(x,\xi)=\frac{2^{p+q-2}\Gamma(p/2)\Gamma(q/2)}{\pi^{n/2}\Gamma(p)\Gamma(q)}\cdot|x_0-\Xi|^{-2k}\cdot F_A\left(k,\frac{p}{2},\frac{q}{2},p,q,-\frac{4\Xi_{n-1}x_{n-1}}{|x-\Xi|^2},0\right)$$

and also

$$\lim_{\xi\to\Xi}\hat{E}(x_0,\xi)=\frac{2^{p+q-2}\Gamma(p/2)\Gamma(q/2)}{\pi^{n/2}\Gamma(p)\Gamma(q)}|x_0-\Xi|^{-2k}\cdot F_A\left(k,\frac{p}{2},\frac{q}{2},p,q,\frac{-4\Xi_{n-1}x_{0_{n-1}}}{|x_0-\Xi|^2},0\right).$$

Thus we consider $\lim_{\xi\to\Xi}\hat{E}(x,\xi)/\hat{E}(x_0,\xi)$.

It is obvious from the integral representation for $\hat{E}_{p,q}(x,\xi)$ that the denominator and numerator are both nonzero and so this is equal to

$$\left|\frac{x-\Xi}{x_0-\Xi}\right|^{-2k}\cdot\frac{F_A\left(k,\frac{p}{2},\frac{q}{2},p,q,-\frac{4\Xi_{n-1}x_{n-1}}{|x-\Xi|^2},0\right)}{F_A\left(k,\frac{p}{2},\frac{q}{2},p,q,-\frac{4\Xi_{n-1}x_{0_{n-1}}}{|x_0-\Xi|^2},0\right)}=K(x,\Xi). \tag{3.2}$$

Furthermore, if the $(n-1)$th coordinate of Ξ, Ξ_{n-1} is zero then it is clear that

$$\lim_{\xi\to\Xi}\frac{E(x,\xi)}{E(x_0,\xi)}=\left|\frac{x-\Xi}{x_0-\Xi}\right|^{-2k}=K(x,\Xi). \tag{3.3}$$

In the second case, $|\xi|\to\infty$, and by using (2.1), (3.1) a calculation shows that

$$\lim_{|\xi|\to\infty}\frac{E(x,\xi)}{E(x_0,\xi)}=1. \tag{3.4}$$

Now

$$\lim_{|\xi|\to\infty}\left[\frac{\frac{|x-\xi|^2}{|\xi|^2}+\frac{4x_{n-1}\xi_{n-1}}{|\xi|^2}\sin^2\frac{\theta_1}{2}+\frac{4x_n\xi_n}{|\xi|^2}\sin^2\frac{\theta_2}{2}}{\frac{|x_0-\xi|^2}{|\xi|^2}+\frac{4x_{0_{n-1}}\xi_{n-1}}{|\xi|^2}\sin^2\frac{\theta_1}{2}+\frac{4x_n\xi_n}{|\xi|^2}\sin^2\frac{\theta_2}{2}}\right]^{-k}=\lim_{|\xi|\to\infty}\left[\frac{\sigma(x,\xi)}{\sigma(x_0,\xi)}\right]^{-2k}\to 1.$$

This limit is uniform in θ_1, θ_2 and it leads us to the result (3.4).

(b) The case where p and $q<1$. In this case we can use the correspondence principle; i.e., a function u is a solution of $L_{2-p,2-q}[u]=0$ if and only if $v=x_{n-1}^{1-p}\cdot x_n^{1-q}\cdot u$ is a solution of $L_{p,q}[v]=0$.

Note that the transformation is order preserving and so it preserves minimality, etc.

Using (3.1) we take $K(x,\Xi)$ as u and we get K^1 as v where

$$K^1(x,\Xi)=x_{n-1}^{1-p}\cdot x_n^{1-q}K(x,\Xi).$$

We replace p and q by $2-p$, $2-q$ in (3.1):

$$x_{0_{n-1}}^{1-p}\cdot x_{0_n}^{1-q}\cdot K^1=\begin{cases}x_{n-1}^{1-p}\cdot x_n^{1-q}\cdot K(x,\Xi), & \Xi_{n-1}\neq 0, \quad (3.2')\\ x_{n-1}^{1-p}\cdot x_n^{1-q}\cdot\left|\frac{x-\Xi}{x_0-\Xi}\right|^{p+q-2-n}, & \Xi_{n-1}=\Xi_n=0, \quad (3.3')\\ x_{n-1}^{1-p}\cdot x_n^{1-q}\cdot 1, & |\xi|\to\infty \quad (3.4')\end{cases}$$

where $K(x,\Xi)$ is defined in (3.2) with the parameters $2-p$, $2-q$.

(c) The case where p (or q) <1 and q (or p) >1. We can use the correspondence principle in the following form: a function u is a solution of $L_{2-p,q}[u]=0$ (or $L_{p,2-q}[u]=0$) if and only if $v=x_{n-1}^{1-p}\cdot u$ (or $v=x_n^{1-q}\cdot u$) is a solution of $L_{p,q}[v]=0$. Using the same argument as we did in case (b), we can find

$$x_{0_{n-1}}^{1-p}\cdot K^2=\begin{cases}x_{n-1}^{1-p}K(x,\Xi)[\text{or } x_n^{1-q}K(x,\Xi)], & \Xi_{n-1}\neq 0, \quad (3.2'')\\ x_{n-1}^{1-p}\left|\frac{x-\Xi}{x_0-\Xi}\right|^{p-q-n}\left(\text{or } x_n^{1-q}\left|\frac{x-\Xi}{x_0-\Xi}\right|^{q-p-n}\right), & \Xi_{n-1}=\Xi_n=0, \quad (3.3'')\\ x_{n-1}^{1-p}\cdot 1(\text{or } x_n^{1-q}\cdot 1), & |\xi|\to\Xi=\infty, \quad (3.4'')\end{cases}$$

where $K(x,\Xi)$ is also defined in (3.2) with the parameters $2-p$, q (or p, $2-q$).

4. The integral representation of the positive solutions of (1.1). Using the Harnack theorem we readily see that $K(x,\Xi)$ is a locally uniform limit of harmonic functions

in x and so in turn is harmonic [17]. Thus if μ is a finite positive measure on ∂Q, and we see that

$$h(x)=\int_{\partial Q} K(x,\Xi)\,d\mu(\Xi) \tag{4.1}$$

is a positive harmonic function on Q.

Letting Φ be a collection of functions $E(x,\xi)/E(x_0,\xi)$ and $\Omega=Q$, we see that $Q\cup\partial Q\cup\{\infty\}$ satisfies the role of $\hat{\Omega}$ in [14] and $\hat{\Omega}$ is the Constantinescu-Cornea compactification of Q. $\partial Q\cup\{\infty\}$ is the Martin boundary and also $\hat{\Omega}$ is called Martin space. Following Martin, as described in [14] or in the original article by Martin [13], any positive harmonic function u on Q may be expressed in the form

$$u(x)=\int_{\partial Q} K(x,\Xi)\,d\mu(\Xi) \tag{4.2}$$

where μ is a finite positive Radon measure. Moreover, if

$$\partial\tilde{Q}=\{\Xi|K(x,\Xi)\text{ is a minimal harmonic function}\}$$

then there is a measure $\tilde{\mu}$ on $\partial\tilde{Q}$ that also gives u and among measures supported by $\partial\tilde{Q}$ it is unique. This is because the cone of all positive harmonic functions on Q that are one at x_0 is a compact Choquet simplex and so every positive harmonic function is the barycenter of a unique measure on the extreme points and these in turn are just minimal harmonic functions.

The following definition is Martin's [13]: A function positive and harmonic in a given domain is called minimal— for this domain—if it dominates no positive harmonic function on the entire domain except for its own constant submultiples.

We wish to identify the minimal harmonic functions. It is evident that $\tilde{\partial}Q$ is nonempty—indeed, it must contain at least one $\Xi\neq\infty$ that does have $\Xi_n>0$; otherwise we note that in the case $p,q<1$ every positive harmonic function on Q (being an integral combination of such functions) would be zero on the boundary hyperplane $\Xi_n>0$. By a scaling argument if $K(x,\Xi)$ is minimal so is $K(x,\lambda\Xi)$, etc. Thus every $\Xi\in\partial Q$ with $\Xi_n>0$ (or $\Xi_{n-1}>0$) is in $\tilde{\partial}Q$.

PROPOSITION. *Let us suppose $p,q<1$, $u>0$ is a solution of* (1.1) *and u tends to zero on the ∂Q. Then u is proportional to $x_{n-1}^{1-p}\cdot x_n^{1-q}$ and so $x_{n-1}^{1-p}x_n^{1-q}$ is minimal.*

Proof. It is already known [6] that

$$u(x)=x_{n-1}^{1-p}\cdot x_n^{1-q}\int_{Q_r}\xi_{n-1}\xi_n J(x,\xi)g(\xi)\,dQ_r \tag{4.3}$$

is the unique solution of (1.1) in $Q\cap B_r^+$ taking on the boundary values g

$$\lim_{\substack{x\to\xi\\ \xi\in\partial B_r^+}} u(x)=g(\xi),\qquad x\in B_r^+,\quad \xi\in\partial B_r^+$$

where

$$J(x,\xi)=\frac{2(r^2-|x|^2)\Gamma\left(\dfrac{n+4-p-q}{2}\right)}{r\cdot\pi^{n/2}\cdot\Gamma\left(\dfrac{2-p}{2}\right)\Gamma\left(\dfrac{2-q}{2}\right)}\int_0^{\pi}\int_0^{\pi}\sigma^{p+q-n-4}\sin^{1-p}\theta_1\cdot\sin^{1-q}\theta_2\,d\theta_1\,d\theta_2,$$

$$B_r^+=\{x\colon |x|<r,\ x=(x_1,\cdots,x_n)\in\mathbb{R}^n,\ x_{n-1},x_n>0\},$$

$$Q_r=\{x\colon |x|=r,\ x=(x_1,\cdots,x_n)\in\mathbb{R}^n,\ x_{n-1},x_n>0\}.$$

Now we need to show that for each $x', x'' \in B_r^+$

$$\lim_{|\Xi|=r\to\infty} \frac{u(x')/x_{n-1}'^{1-p} \cdot x_n'^{1-q}}{u(x'')/x_{n-1}''^{1-p} \cdot x_n''^{1-q}} \to 1. \tag{4.4}$$

It is obvious that

$$\lim_{r\to\infty} \frac{(r^2-|x'|^2)/r}{(r^2-|x''|^2)/r} \to 1, \qquad \lim_{r\to\infty} \frac{\sigma^{p+q-n-4}(x')}{\sigma^{p+q-n-4}(x'')} \to 1.$$

The last limit is uniform in θ_1, θ_2. Thus,

$$\lim_{|\Xi|\to\infty} \frac{\int_0^\pi \int_0^\pi \sigma^{p+q-n-4}(x') \sin^{1-p}\theta_1 \sin^{1-q}\theta_2 \, d\theta_1 \, d\theta_2}{\int_0^\pi \int_0^\pi \sigma^{p+q-n-4}(x'') \sin^{1-p}\theta_1 \sin^{1-q}\theta_2 \, d\theta_1 \, d\theta_2} \to 1.$$

From this result, we can see that (4.4) is satisfied.

In this proof, $x', x'' \in B_r^+$ are any points in B_r^+, so (4.4) is true in whole B_r^+. Then we can write the following result:

$$\frac{u(x')/x_{n-1}'^{1-p} x_n'^{1-q}}{(x'')/x_{n-1}''^{1-p} \cdot x_n''^{1-q}} = 1. \tag{4.5}$$

This leads to the following result:

$$u(x) = \alpha \cdot x_{n-1}^{1-p} \cdot x_n^{1-q}. \tag{4.6}$$

In (4.5) and (4.6) x, x', x'' are any points in B_r^+. If $u(x)$ is a solution of (1.1) then the function

$$u\left(\frac{x_1}{|x|^2}, \ldots, \frac{x_n}{|x|^2}\right) \cdot |x|^{2-n-p-q}$$

is also a solution of the same equation.

This transformation preserves the property of being minimal (and other properties coming from order structure). It is a Kelvin transformation [12], [18]. Using this inversion we deduce that $K(x, \Xi)$ is minimal for all Ξ in $\Xi_{n-1} = \Xi_n = 0$, and hence for all Ξ on the Martin boundary when $p, q < 1$. In this case we therefore have that any positive solution u to $L_{p,q}(u) = 0$ is of the form

$$u(x) = \int_{\partial Q \cup \{\infty\}} K(x, \Xi) \, d\mu$$

where μ is a uniquely determined Radon measure.

We can transfer to the other cases for p, q by using the correspondence principle. The appropriate formulae for $K(x, \Xi)$ are given by

(a) (3.2), (3.3), (3.4) when $p, q > 1$,

(b) (3.2′), (3.3′), (3.4′) when $p, q < 1$,

(c) (3.2″), (3.3″), (3.4″) when $p < 1$, $q > 1$ (or $p > 1$, $q < 1$).

Using Theorem 3 of [6] the kernel gives a unique integral representation for every positive solution of (1.1).

The cases where $p = 1$ (or $q = 1$) and $p = q = 1$ can be treated easily.

Acknowledgment. The author expresses his gratitude to Professor Terry J. Lyons, Department of Mathematics, University of Edinburgh, for suggesting the topic and for his guidance throughout the development of this work.

REFERENCES

[1] R. P. GILBERT, *Integral operator methods in bi-axially symmetric potential theory*, Contrib. to Differential Equations, 2 (1963), pp. 441-456.

[2] ———, *Function Theoretic Methods in Partial Differential Equations*, Academic Press, New York, 1963.

[3] R. J. WEINACHT, *Fundamental solutions for a class of equations with several singular coefficients*, J. Austral. Math. Soc., 8 (1968), pp. 575-583.

[4] M. B. KAPILEVICH, *Mean value theorems for solutions of singular elliptic differential equations*, Izv. Vyssh. Uchebn. Zaved. Mat., 19 (1960), pp. 114-125.

[5] A. O. CELEBI, *Some properties for solutions of* $\sum_{i=1}^{n}((\partial^2/\partial x_i^2)+(k_i/x_i)(\partial/\partial x_i))u=0$, Comm. Fac. Sci. Univ. Ankara Sér. A, 19 (1970), pp. 31-39.

[6] N. S. HALL, D. W. QUINN, AND R. J. WEINACHT, *Poisson integral formulas in generalized bi-axially symmetric potential theory*, SIAM J. Math. Anal., 5 (1974), pp. 111-118.

[7] D. W. QUINN AND R. J. WEINACHT, *Boundary value problems in generalised bi-axially symmetric potential theory*, J. Differential Equations, 21 (1976), pp. 113-133.

[8] A. HUBER, *On the uniqueness of generalized axially symmetric potentials*, Annals of Math., 60 (1954), pp. 351-358.

[9] ———, *Some results on generalized axially symmetric potentials*, in Proc. Conference on Differential Equations, University of Maryland, College Park, MD, 1956, pp. 147-155.

[10] J. B. DIAZ AND A. WEINSTEIN, *On the fundamental solutions of a singular Beltrami operator*, Studies in Mathematics and Mechanics presented to R. Von Mises, Academic Press, New York, 1954, pp. 97-102.

[11] R. J. WEINACHT, *Fundamental solutions for a class of singular equations*, Contrib. to Differential Equations, 3 (1964), pp. 43-55.

[12] B. BRELOT-COLLIN AND M. BRELOT, *Representation integrale de solutions positives de l'equation* $L_k[u]=\sum_{i=1}^{n}(\partial^2u/\partial x_i^2)+(k/x_n)(\partial u/\partial x_n)=0$ (k constante réelle dans demi-space $E(x_n>0)$ de $\mathbb{R}^n$), Bull. Acad. Royale des Sciences de Bruxelles, 58 (1972), pp. 317-326.

[13] R. S. MARTIN, *Minimal positive harmonic functions*, Trans. Amer. Math. Soc., 49 (1941), pp. 137-172.

[14] M. BRELOT, *On Topologies and Boundaries in Potential Theory*, Lecture Notes in Mathematics 175, Tata Institute of Fundamental Research, Bombay, India, 1971.

[15] ———, *Topology of R. S. Martin and Green Lines*, Lectures on Functions of a Complex Variable, The University of Michigan Press, Ann Arbor, MI, 1955, pp. 105-121.

[16] P. APPEL AND J. KAMPÉ DE FÉRIET, *Fonctions Hypergéometriques et hypersphériques, polynomes d'hermite*, Gauthier-Villars, Paris, 1926, p. 115.

[17] M. BRELOT, *Lectures on potential theory*, Collect. Math. du Tata Institute, 19 (1960); réedité 1967.

[18] A. O. CELEBI, *On the generalized Tricomi's equation*, Comm. Fac. Sci. Univ. Ankara Sér. A, 17 (1968), pp. 1-31.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

008

EXISTENCE OF SOLUTIONS TO THE STOMMEL–CHARNEY MODEL OF THE GULF STREAM*

V. BARCILON†, P. CONSTANTIN‡, AND E. S. TITI§

Abstract. The existence of weak solutions to the equations proposed by Stommel [*Trans. Amer. Geophys. Union*, 29 (1948), pp. 202–206] and Charney [*Proc. Nat. Acad. Sci. U.S.A.*, 41 (1955), pp. 731–740] as a model of the Gulf Stream are established by means of the method of artificial viscosity.

Key words. Navier–Stokes equations, artificial viscosity, ocean circulation

AMS(MOS) subject classifications. 35Q10, 76U05

1. Introduction. In this paper, we examine the mathematical properties of an equation arising in the theory of ocean circulation. In order to understand the role of this problem in oceanography, a brief review of the subject is necessary. The first successful attempt to provide a mathematical description of the mid-latitude ocean currents was made by Stommel [13] in 1948. He showed conclusively that a Gulf Stream-like intensification on the western side of an ocean basin could be explained by the so-called β-effect. This is the geophysical terminology for the latitudinal variation of the normal component of the earth's rotation. Aside from this variable Coriolis force, the other forces which entered into Stommel's model were those due to the pressure gradient, the surface winds, and friction. For the sake of simplicity, this last force was taken to be proportional to the velocity fields. All the effects of density stratification were neglected by making the assumption that the ocean was homogeneous. Finally, by working with vertical averages, Stommel essentially treated the ocean circulation as a two-dimensional horizontal motion. Somewhat surprisingly, Stommel's ad hoc, linear model was shown later to provide an accurate description of an actual experimental setup [12]. In particular, it is now well known that the Rayleigh friction terms included in Stommel's model arise quite naturally from a consideration of frictional effects near the ocean bottom.

The subsequent work in the field has attempted to overcome the two oversimplifications of Stommel's model, namely, to take into account inertial forces that are known to be important in the vicinity of the Gulf Stream and to represent more adequately the complex mixing/dissipative processes. These two broad generalizations were initiated by Charney [5] and Munk and Carrier [10], respectively. Although these early papers make use of analytical techniques, the bulk of the recent work in the field has relied on numerical techniques to study the nonlinear problems arising in the mathematical formulation of wind-driven ocean circulation models. The papers by Bryan [4] and Veronis [15], [16] fall within this category.

* Received by the editors May 4, 1987; accepted for publication (in revised form) February 22, 1988.

† Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois 60637. The work of this author was supported by Office of Naval Research grant N00014-86-K-0034.

‡ Department of Mathematics, University of Chicago, Chicago, Illinois 60637. The work of this author was supported by National Science Foundation grant DMS-860-2031. This author was a Sloan Research Fellow during the course of this work.

§ Department of Mathematics, University of Chicago, Chicago, Illinois 60637. The work of this author was supported by Department of Energy grant DE-FG02-86ER25020.

In this paper, we study the inertial extension of Stommel's model, namely

$$\varepsilon\Delta\psi+\frac{\partial\psi}{\partial x_1}+R\left(\frac{\partial\psi}{\partial x_1}\frac{\partial\Delta\psi}{\partial x_2}-\frac{\partial\psi}{\partial x_2}\frac{\partial\Delta\psi}{\partial x_1}\right)=f \tag{1.1}$$

in a bounded domain Ω in R^2 with Dirichlet boundary conditions. Just as in Stommel's paper, ψ is the streamfunction of the vertically averaged horizontal velocity fields. Hence, $\Delta\psi$ is the vertically integrated relative planetary vorticity which we will also denote by ω. The right-hand side f represents the driving due to the curl of the applied surface wind stress. Each of the terms on the left-hand side has a physical interpretation: the first represents the loss of vorticity by frictional processes along the ocean bottom, the second the gain in vorticity due to a northern flow (this is the β-effect), and the third the advection of vorticity by the circulation. Thus, the above equation represents the vorticity budget of a column of fluid in the ocean. The parameters R and ε arise as a result of the nondimensionalization of the various fields. Such a nondimensionalization can be carried out in several different ways depending upon which process controlling the vorticity is perceived as being crucial. In the ocean, measurements suggest that the prevailing balance is the so-called Sverdrup balance in which the β-effect is comparable to the wind curl. This is why the corresponding terms in (1.1) have unit coefficients. The other terms in (1.1) represent small departures from this balance. Thus, we consider both R and ε as small parameters. They represent, respectively, the importance of inertial and frictional effects relative to the β-effect. The interested reader is referred to Pedlosky [11] for a very thorough discussion of the derivation of this equation.

The main result of this paper is the proof of the existence of weak solutions to (1.1). These weak solutions are obtained as limits of solutions of auxiliary equations with artificial viscosity and artificial boundary conditions. Physically, this auxiliary problem corresponds to the addition of side-wall friction and stress-free boundary conditions. We derive uniform L^∞ bounds for the solutions of the auxiliary equations. This procedure cannot yield classical solutions because the artificial boundary conditions give rise to boundary layers in the classical limit. Nevertheless, we expect the solutions to be classical in certain parameter ranges.

The same method was used by Yudovitch [17] in his paper dealing with the time-dependent two-dimensional Euler equations. This method cannot be extended to the time-independent Euler equations. Actually, it is known [9] that the time-independent Euler equations have no solutions for two- and three-dimensional axisymmetric domains and generic driving forces. However, because of the presence of the bottom friction term $\varepsilon\Delta\psi$, the method can be used for (1.1).

Once the estimates for the stationary problem (1.1) are understood, solutions for the time-dependent version of this problem

$$\frac{\partial\Delta\psi}{\partial t}+\varepsilon\Delta\psi+\frac{\partial\psi}{\partial x_1}+R\left(\frac{\partial\psi}{\partial x_1}\frac{\partial\Delta\psi}{\partial x_2}-\frac{\partial\psi}{\partial x_2}\frac{\partial\Delta\psi}{\partial x_1}\right)=f \tag{1.2}$$

with given initial conditions on $\Delta\psi$ can be obtained by applying the method of Yudovitch in a straightforward manner. The β-effect term does not create any serious difficulties. We can show that (1.2) has unique global solutions.

The problem of describing the set of stationary solutions is still open. We expect the solution to be unique only when ε is large compared to R.

2. Preliminaries. Let Ω be an open bounded set of R^2 with sufficiently smooth boundary $\partial\Omega$. We consider the system

$$\begin{aligned} &\varepsilon\omega + \frac{\partial\psi}{\partial x_1} + R\left(\frac{\partial\psi}{\partial x_1}\frac{\partial\omega}{\partial x_2} - \frac{\partial\psi}{\partial x_2}\frac{\partial\omega}{\partial x_1}\right) = f \quad \text{in } \Omega, \\ &\Delta\psi = \omega \qquad\qquad\qquad\qquad\qquad\qquad \text{in } \Omega, \\ &\psi = 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $f(x)$, $\varepsilon > 0$ and $R \geqq 0$ are given.

In this paper we study the existence of weak solutions of (2.1).

We denote by $H^s(\Omega)$ the usual Sobolev spaces of order s and by $H_0^1(\Omega)$ the closure of $C_0^\infty(\Omega)$ in the $H^1(\Omega)$ norm. We define

$$A = -\Delta$$

to be the negative Laplacian with domain $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$. It is well known that A^{-1} is a compact linear self-adjoint positive operator in $L^2(\Omega)$ (cf. [2], [8]). The spectrum of A consists of an infinite sequence $0 < \lambda_1 < \lambda_2 \leqq \cdots$ of eigenvalues counted according to their multiplicities; $\lambda_n \to \infty$ as $n \to \infty$; the eigenfunctions $\{w_n\}$ provide an orthonormal basis in $L^2(\Omega)$. Finally, there exists a scale invariant constant c_0 such that $|\Omega| = c_0\lambda_1^{-1}$ where $|\Omega|$ denotes the area of Ω. The scalar product and norm in $L^2(\Omega)$ are denoted by $(\cdot,\cdot)$ and $|\cdot|$, respectively. A scalar product in $H_0^1(\Omega)$, in view of Poincaré's inequality, is

$$((u, v)) = \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx \quad \forall u, v \in H_0^1(\Omega),$$

and the corresponding norm is denoted by $\|\cdot\|$. The norm in $L^p(\Omega)$ for $1 \leqq p \leqq \infty$ is denoted by $\|\cdot\|_p$. It is known (cf. [2]) that on $D(A)$ the $H^2(\Omega)$ norm is equivalent to the $|A\cdot|$ norm, i.e., there exists a constant $c_1 > 0$ such that

$$c_1^{-1}|Au| \leqq \|u\|_{H^2(\Omega)} \leqq c_1|Au| \quad \forall u \in D(A).$$

Moreover, $D(A^{1/2}) = H_0^1(\Omega)$ and $\|\cdot\| = |A^{1/2}\cdot|$.

The bilinear bounded operator $J: H^1(\Omega) \times H^1(\Omega) \to L^1(\Omega)$ is defined as

$$J(\psi, \omega) = \frac{\partial\psi}{\partial x_1}\frac{\partial\omega}{\partial x_2} - \frac{\partial\psi}{\partial x_2}\frac{\partial\omega}{\partial x_1} \quad \forall \psi, \omega \in H^1(\Omega). \tag{2.2}$$

We need the following inequalities that correspond to various continuity properties of the operator J.

PROPOSITION 2.1. *Let $s_1, s_2, s_3 \geqq 0$ satisfy*

$$s_1 + s_2 + s_3 \geqq 1 \quad \text{if } s_i \neq 1 \quad \text{for all } i = 1, 2, 3$$

and

$$s_1 + s_2 + s_3 > 1 \quad \text{if } s_i = 1 \quad \text{for some } i = 1, 2, 3.$$

Then, there exists a scale invariant constant $c(s_1, s_2, s_3)$ such that

$$|(J(\psi, \omega), v)| \leqq c(s_1, s_2, s_3)\lambda_1^{(s_1+s_2+s_3)/2}\|\psi\|_{H^{s_1+1}(\Omega)} \cdot \|\omega\|_{H^{s_2+1}(\Omega)} \cdot \|v\|_{H^{s_3}(\Omega)}. \tag{2.3}$$

The reader is referred to [14] and [6] for the idea of the proof. We also note that

$$(J(\psi, \omega), v) = -(J(\omega, \psi), v) \tag{2.4}$$

for every $\psi \in H^{s_1+1}(\Omega)$, $\omega \in H^{s_2+1}(\Omega)$, and $v \in H^{s_3}(\Omega)$.

PROPOSITION 2.2. *For every* $\psi \in H^2(\Omega)$, $\omega \in H^1(\Omega)$ *and* $v \in H_0^1(\Omega)$ *we have*

$$(J(\psi, \omega), v) = -(J(\psi, v), \omega). \tag{2.5}$$

Proof. Because of Proposition 2.1 it is sufficient to show that (2.5) holds for $v \in C_0^\infty(\Omega)$ and $\psi, \omega \in C^\infty(\Omega)$. Let $\mathbf{u}$ be the vector $(-\partial\psi/\partial x_2, \partial\psi/\partial x_1)$; then we can write

$$(J(\psi, \omega), v) = \int_\Omega \nabla \cdot (\omega(x)\mathbf{u}(x))v(x)\, dx$$

$$= -\int_\Omega (\mathbf{u}(x) \cdot \nabla v(x))\omega(x)\, dx.$$

The following corollary is an immediate consequence of the above proposition.

COROLLARY 2.3.

$$(J(\psi, \omega), \omega) = 0 \quad \forall \psi \in H^2(\Omega), \quad \omega \in H_0^1(\Omega), \tag{2.6}$$

$$(J(\psi, \omega), \omega^p) = 0 \quad \forall \psi \in H^2(\Omega), \quad \omega \in D(A), \quad p = 1, 2, \cdots. \tag{2.7}$$

Proof. Equation (2.6) is a direct consequence of (2.5). Since $H^2(\Omega)$ is a Banach algebra (cf. [1]) we can easily verify that $\omega^p \in D(A)$, for $p = 1, 2, \cdots$, whenever $\omega \in D(A)$. We can then use (2.5) to deduce that

$$(J(\psi, \omega), \omega^p) = -(J(\psi, \omega^p), \omega),$$

while from direct computations

$$(J(\psi, \omega^p), \omega) = -p(J(\psi, \omega), \omega), \qquad p = 1, 2, \cdots,$$

which verifies (2.7), in view of (2.6). Let us note that (2.6) also holds for $\psi \in C^1(\bar{\Omega})$, $\omega \in C^1(\bar{\Omega})$, and $\nabla\psi$ normal to $\partial\Omega$.

3. A stationary problem with artificial viscosity. In this section we consider a singular perturbation of the stationary problem (2.1) obtained by adding an artificial viscosity term $-\nu A\omega$ and imposing homogeneous Dirichlet boundary conditions on ω. The artificial viscosity equation provides an approximate solution. Uniform bounds, independent of the viscosity ν will enable us to pass to the weak limit. In this process the boundary condition on ω is lost. The auxiliary problem is:

$$-\nu\Delta\omega + \varepsilon\omega + \frac{\partial\psi}{\partial x_1} + RJ(\psi, \omega) = f \quad \text{in } \Omega, \tag{3.1a}$$

$$-\Delta\psi + \omega = 0 \quad \text{in } \Omega, \tag{3.1b}$$

$$\psi = 0, \quad \omega = 0 \quad \text{on } \partial\Omega, \tag{3.1c}$$

where ν is given such that $0 < \nu < \varepsilon^3/8$ and $f \in L^\infty(\Omega)$.

3.1. Existence of solutions of the perturbed problem. We restate the problem: Find $(\omega, \psi) \in D(A) \times D(A)$ such that

$$\nu A\omega + \varepsilon\omega + \frac{\partial\psi}{\partial x_1} + RJ(\psi, \omega) = f, \tag{3.2a}$$

$$A\psi + \omega = 0. \tag{3.2b}$$

It is obvious that every solution (ω, ψ) of (3.2) is a solution to (3.1) in the distribution sense.

The following a priori estimates of the solutions of (3.2) will be needed.

LEMMA 3.1. *Let* $(\omega, \psi) \in D(A) \times D(A)$ *be a solution of* (3.2). *Then*

$$\nu \|\omega\|^2 + \frac{\varepsilon}{4} |\omega|^2 \leqq \frac{|f|^2}{\varepsilon} K(\varepsilon, \lambda_1), \tag{3.3}$$

$$\|\psi\|^2 \leqq 2|f|^2 K(\varepsilon, \lambda_1), \tag{3.4}$$

$$|A\psi|^2 \leqq \frac{4|f|^2}{\varepsilon^2} K(\varepsilon, \lambda_1), \tag{3.5}$$

and

$$\nu |A\omega| \leqq 2|f| + 2\sqrt{2}|f| K^{1/2}(\varepsilon, \lambda_1) \left[1 + \frac{R^2 c^2(\tfrac{1}{2}, \tfrac{1}{2}, 0) c_2^2}{\nu^{3/2} \varepsilon^{1/2}} |f|^2 K(\varepsilon, \lambda_1) \right], \tag{3.6}$$

where

$$K(\varepsilon, \lambda_1) = \left(1 + \frac{1}{\varepsilon^2 \lambda_1} \right).$$

Proof. Let $(\omega, \psi) \in D(A) \times D(A)$ be a solution to (3.2). Taking the scalar product of (3.2a) with ψ we get

$$\nu(A\omega, \psi) + \varepsilon(\omega, \psi) + \left(\frac{\partial \psi}{\partial x_1}, \psi \right) + R(J(\psi, \omega), \psi) = (f, \psi). \tag{3.7}$$

In view of (2.4) and (2.5), and of the fact that $(\partial\psi/\partial x_1, \psi) = 0$, (3.7) reduces to

$$\nu(A\omega, \psi) + \varepsilon(\omega, \psi) = (f, \psi),$$

or, because of (3.2b), to

$$\varepsilon \|\psi\|^2 \leqq \nu |\omega|^2 + |f||\psi|.$$

Using first Poincaré's inequality we can write

$$\varepsilon \|\psi\|^2 \leqq \nu |\omega|^2 + |f| \lambda_1^{-1/2} \|\psi\|.$$

Then by Young's inequality, we can write

$$\varepsilon \|\psi\|^2 \leqq \nu |\omega|^2 + \frac{|f|^2}{2\varepsilon\lambda_1} + \frac{\varepsilon}{2} \|\psi\|^2.$$

Finally, since $0 \leqq \nu \leqq \varepsilon^3/8$, we deduce

$$\|\psi\|^2 \leqq (2\nu/\varepsilon)|\omega|^2 + \frac{|f|^2}{\varepsilon^2 \lambda_1} \leqq \frac{\varepsilon^2}{4} |\omega|^2 + \frac{|f|^2}{\varepsilon^2 \lambda_1}.$$

If we take the scalar product of (3.2a) with ω and use (2.6) we obtain

$$\begin{aligned} \nu \|\omega\|^2 + \varepsilon |\omega|^2 &= -(\partial\psi/\partial x_1, \omega) + (f, \omega) \\ &\leqq \|\psi\| |\omega| + |f||\omega|, \end{aligned}$$

and by Young's inequality

$$\nu \|\omega\|^2 + \frac{\varepsilon}{2} |\omega|^2 \leqq \frac{\|\psi\|^2}{\varepsilon} + \frac{|f|^2}{\varepsilon}.$$

From the above inequality and (3.8) we deduce (3.3). From (3.3) and (3.8) we can easily obtain (3.4). Finally, (3.2b) and (3.3) directly imply (3.5).

To establish (3.6) we take the scalar product of (3.2a) with $A\omega$, from which we see that

$$\nu|A\omega|^2+\varepsilon\|\omega\|^2\leqq|(\partial\psi/\partial x_1, A\omega)|+R|(J(\psi,\omega), A\omega)|+|(f, A\omega)|,$$

$$\leqq\|\psi\|\,|A\omega|+R|(J(\psi,\omega), A\omega)|+|(f, A\omega)|.$$

By using (2.3) for $s_1=s_2=\frac{1}{2}$ and $s_3=0$, we get

$$\nu|A\omega|^2+\varepsilon\|\omega\|^2\leqq(\|\psi\|+Rc(\tfrac{1}{2},\tfrac{1}{2},0)\|\psi\|_{H^{3/2}(\Omega)}\|\omega\|_{H^{3/2}(\Omega)}+|f|)|A\omega|,$$

and by an interpolation inequality, we get

$$\nu|A\omega|^2+\varepsilon\|\omega\|^2\leqq(\|\psi\|+Rc(\tfrac{1}{2},\tfrac{1}{2},0)c_2\|\psi\|^{1/2}\|A\psi\|^{1/2}\|\omega\|^{1/2}|A\omega|^{1/2}+|f|)|A\omega|;$$

hence

$$\frac{\nu}{2}|A\omega|^2+\varepsilon\|\omega\|^2\leqq\left(\|\psi\|+\frac{R^2}{2\nu}c^2(\tfrac{1}{2},\tfrac{1}{2},0)c_2^2\|\psi\|\,|A\psi|\,\|\omega\|+|f|\right)|A\omega|.$$

Making use of (3.3), (3.4), and (3.5), we obtain (3.6).

LEMMA 3.2. *There exists at least one solution to problem* (3.2). *Moreover, every solution of* (3.2) *satisfies* (3.3)–(3.6).

Proof. One approach is to use the Galerkin approximation method based on the eigenfunctions of the operator A as in Constantin and Foias [6] and Temam [14] to show the existence of a solution to (3.2). The crucial point in this approach is to establish similar a priori estimates to the ones in (3.5)–(3.6) for the approximate solution.

Another approach is by the Leray–Schauder degree theory. Indeed, problem (3.2) is equivalent to

$$\begin{pmatrix}\omega\\ \psi\end{pmatrix}+\mathbf{K}(\psi,\omega)=\begin{pmatrix}A^{-1}f\\ 0\end{pmatrix} \tag{3.8}$$

where

$$\mathbf{K}(\psi,\omega)=\begin{pmatrix}\nu^{-1}A^{-1}[\varepsilon\omega+\partial\psi/\partial x_1+RJ(\psi,\omega)]\\ A^{-1}\omega\end{pmatrix}. \tag{3.9}$$

By Rellich's lemma and (2.3), we can show that the nonlinear mapping

$$\mathbf{K}: D(A)\times D(A)\to L^2(\Omega)\times L^2(\Omega)$$

is compact. Therefore, making use of the a priori estimates (3.5)–(3.6), we can conclude that (3.8) has at least one solution [3].

Remark 3.3. If $f\in C^\infty(\bar\Omega)$, then every solution $(\omega,\psi)\in D(A)\times D(A)$ of (3.2) satisfies (3.1) in the classical sense, namely $\psi, \omega\in C^\infty(\bar\Omega)$.

3.2. Uniform L^∞ bounds for the artificial viscosity problem. In (3.3), (3.4), and (3.5) we gave estimates for $|\omega|$, $\|\psi\|$, and $|A\psi|$, which were independent of ν. In this section we shall derive uniform (i.e., independent of ν) L^∞ estimates for every solution of (3.2) and for $0<\nu<\varepsilon^3/8$. We recall first the following known result in potential theory (see, e.g., [7]).

THEOREM 3.4. *Let $G(x, y)$ denote the Green function of the Laplacian operator in the domain with Dirichlet boundary conditions. Then, there exists two scale invariant constants c_3, c_4 such that*

$$|G(x, y)| \leqq c_3(1+|\log(\lambda_1^{1/2}|x-y|)|) \quad \forall x, y \in \Omega, \quad x \neq y, \tag{3.10}$$

and

$$\left|\frac{\partial G(x, y)}{\partial x_k}\right| \leqq \frac{c_4}{|x-y|} \quad \forall k = 1, 2, \quad x, y \in \Omega, \quad x \neq y. \tag{3.11}$$

Moreover, since

$$\phi(x) = \int_\Omega G(x, y)\Delta\phi(y)\, dy,$$

it follows that

$$\frac{\partial \phi(x)}{\partial x_k} = \int_\Omega \frac{\partial G(x, y)}{\partial x_k} \Delta\phi(y)\, dy. \tag{3.12}$$

THEOREM 3.5 (Uniform L^∞ bounds). *Let $(\omega, \psi) \in D(A) \times D(A)$ be a solution of* (3.2) (*or equivalently a solution of* (3.1)). *Then*

$$\|\omega\|_\infty \leqq \frac{2}{\varepsilon}\|f\|_\infty + \frac{64\pi^2 c_4^2 c_0^{1/2}}{\varepsilon^3 \lambda_1^{1/2}} |f| K^{1/2}(\varepsilon, \lambda_1) \tag{3.13}$$

and

$$\|\nabla\psi\|_\infty \leqq 4\sqrt{2\pi} c_4 c_0^{1/4} \lambda_1^{-1/4} \sqrt{\|\omega\|_\infty |\omega|}. \tag{3.14}$$

Note that (3.3) *provides a uniform upper bound for* $|\omega|$.

Proof. Let $\delta > 0$. We denote by $B_\delta(z)$ the ball in R^2 that is centered at z with radius δ. From (3.1a) and (3.13) we have

$$\nabla\psi(x) = \int_\Omega \nabla_x G(x, y)\omega(y)\, dy,$$

which, because of (3.11), implies

$$|\nabla\psi(x)| \leqq 2c_4 \int_\Omega \frac{|\omega(y)|}{|x-y|}\, dy,$$

or

$$|\nabla\psi(x)| \leqq 2c_4 \left\{ \int_{\Omega \cap B_\delta(x)} \frac{|\omega(y)|}{|x-y|}\, dy + \int_{\Omega \setminus B_\delta(x)} \frac{|\omega(y)|}{|x-y|}\, dy \right\}.$$

Using Hölder's inequality, we see that for odd integers $p \geqq 3$

$$|\nabla\psi(x)| \leqq 2c_4 \left\{ \|\omega\|_{p+1} \left(\int_{B_\delta(x)} |x-y|^{-(p+1)/p}\, dy \right)^{p/p+1} + \frac{|\omega|}{\delta} |\Omega|^{1/2} \right\},$$

i.e.,

$$|\nabla\psi(x)| \leqq 2c_4 \left\{ \|\omega\|_{p+1} (2\pi)^{p/p+1} \delta^{(p-1)/(p+1)} + \frac{|\omega|}{\delta} |\Omega|^{1/2} \right\}.$$

Therefore

$$\|\nabla\psi\|_\infty \leqq (4\pi c_4)\|\omega\|_{p+1} \delta^{p-1/p+1} + 2c_4 |\Omega|^{1/2} \frac{|\omega|}{\delta}. \tag{3.15}$$

In order to get an estimate for $\|\omega\|_{p+1}$, we multiply (3.1a) by $\omega^p(x)$ and integrate over Ω. This leads to

$$-\nu\int_\Omega \Delta\omega(x)\omega^p(x)\,dx+\varepsilon\int_\Omega \omega^{p+1}(x)\,dx+R(J(\psi,\omega),\omega^p)=-\left(\frac{\partial\psi}{\partial x_1},\omega^p\right)+(f,\omega),$$

or, after integrating by parts and using (2.7), to

$$p\nu\int_\Omega |\nabla\omega(x)|^2\omega^{p-1}(x)\,dx+\varepsilon\int_\Omega \omega^{p+1}(x)\,dx=-\left(\frac{\partial\psi}{\partial x_1},\omega^p\right)+(f,\omega).$$

Using Hölder's inequality as well as the fact that p is an odd integer, we conclude that

$$\varepsilon\|\omega\|_{p+1}^{p+1}\leqq\|\nabla\psi\|_\infty\|\omega\|_p^p+\|f\|_{p+1}\|\omega\|_{p+1}^p,$$

or

$$\varepsilon\|\omega\|_{p+1}^{p+1}\leqq(\|\nabla\psi\|_\infty|\Omega|^{1/p+1}+\|f\|_{p+1})\|\omega\|_{p+1}^p. \tag{3.16}$$

Substituting (3.15) in (3.16) we see that

$$\varepsilon\|\omega\|_{p+1}\leqq\|f\|_{p+1}+(4\pi c_4)|\Omega|^{1/(p+1)}\delta^{(p-1)/(p+1)}\|\omega\|_{p+1}+2c_4|\Omega|^{(p+3)/2(p+1)}\frac{|\omega|}{\delta}. \tag{3.17}$$

If we choose $\delta=(\varepsilon/8\pi c_4)|\Omega|^{-1/(p-1)}$, then (3.17) implies that

$$\|\omega\|_{p+1}\leqq\frac{2}{\varepsilon}\|f\|_{p+1}+\frac{4\pi c_4}{\varepsilon}\left(\frac{8\pi c_4}{\varepsilon}|\Omega|^{1/2}\right)^{p+1/p-1}|\Omega|^{1/p+1}|\omega|. \tag{3.18}$$

Passing to the limit as $p\to\infty$, we conclude that

$$\|\omega\|_\infty\leqq\frac{2}{\varepsilon}\|f\|_\infty+\frac{32\pi^2c_4^2}{\varepsilon^2}|\Omega|^{1/2}|\omega|. \tag{3.19}$$

After replacing $|\Omega|$ in the above expression by $c_0\lambda_1^{-1}$ and using (3.3), we obtain (3.13).

In order to derive (3.14), we observe that (3.15) holds for every $\delta>0$. Therefore

$$\|\nabla\psi\|_\infty\leqq(4\pi c_4)\|\omega\|_\infty\delta+2c_4|\Omega|^{1/2}\frac{|\omega|}{\delta}. \tag{3.20}$$

Minimizing the right-hand side, we deduce that

$$\|\nabla\psi\|_\infty\leqq4\sqrt{2\pi}c_4|\Omega|^{1/4}\sqrt{\|\omega\|_\infty|\omega|},$$

and after replacing $|\Omega|$ by $c_0\lambda_1^{-1}$, we obtain (3.14).

4. Existence of weak solutions. Let $\{\nu_k\}_{k=1}^\infty$ be an arbitrary sequence of real numbers, $\nu_k\in(0,\varepsilon^3/8)$ for $k=1,2,\cdots$, and such that $\nu_k\to0$ as $k\to\infty$. Let (ω_k,ψ_k) be a sequence of solutions to problem (3.1) corresponding to $\nu=\nu_k$ for $k=1,2,\cdots$, respectively. By Lemma 3.2, we know that such sequences exist.

THEOREM 4.1. *There exists a $\psi\in D(A)$ and an $\omega\in L^\infty(\Omega)$ that are weak solutions of the stationary problem* (2.1), *i.e.*,

$$\begin{aligned}&\varepsilon\omega+\frac{\partial\psi}{\partial x_1}+R\nabla\cdot(\omega\mathbf{u})=f && \text{in } \Omega,\\ &\Delta\psi=\omega && \text{in } \Omega,\\ &\psi=0 && \text{on } \partial\Omega,\end{aligned} \tag{4.1}$$

where $\mathbf{u}=(-\partial\psi/\partial x_2, \partial\psi/\partial x_1)$. *Furthermore,* ψ *and* ω *satisfy*

$$|\omega| \leqq \frac{2|f|}{\varepsilon} K^{1/2}(\varepsilon, \lambda_1), \tag{4.2}$$

$$\|\omega\|_\infty \leqq \frac{2}{\varepsilon}\|f\|_\infty + \frac{64\pi^2 c_4^2 c_0^{1/2}}{\varepsilon^3 \lambda_1^{1/2}} |f| K^{1/2}(\varepsilon, \lambda_1), \tag{4.3}$$

$$\|\nabla\psi\|_\infty \leqq 4\sqrt{2\pi} c_4 c_0^{1/4} \lambda_1^{-1/4} \sqrt{\|\omega\|_\infty |\omega|}. \tag{4.4}$$

Proof. Let (ω_k, ψ_k) be as mentioned above. From (3.5) we know that

$$|A\psi_k| \leqq \frac{2|f|}{\varepsilon} K^{1/2}(\varepsilon, \lambda_1).$$

Therefore, there exists a subsequence, say $\{\psi_{k_1}\}$, that converges weakly in $H^2(\Omega)$ to $\psi \in H^2(\Omega)$, and by virtue of Rellich's lemma, ψ_{k_1} converges strongly to $\psi \in H^1(\Omega)$; hence $\psi \in D(A)$. From (3.13) we know that

$$\|\omega_{k_1}\|_\infty \leqq \frac{2}{\varepsilon}\|f\|_\infty + \frac{64\pi^2 c_4^2 c_0^{1/2}}{\varepsilon^3 \lambda_1^{1/2}} |f| K^{1/2}(\varepsilon, \lambda_1),$$

and therefore

$$\|\omega_{k_1}\|_p \leqq |\Omega|^{1/p}\left(\frac{2}{\varepsilon}\|f\|_\infty + \frac{64\pi^2 c_4^2 c_0^{1/2}}{\varepsilon^3 \lambda_1^{1/2}} |f| K^{1/2}(\varepsilon, \lambda_1)\right) \quad \forall p = 1, 2, \cdots. \tag{4.5}$$

By means of (4.5) we can inductively find, for every $p = 2, 3, \cdots$ a subsequence $\{\omega_{k_p}\}$ of $\{\omega_k\}$ that converges weakly in $L^p(\Omega)$ to the same limit $\omega \in L^p(\Omega)$ for every $p = 2, 3, \cdots$. In particular, we have

$$\|\omega\|_p \leqq \liminf_{k_p \to \infty} \|\omega_{k_p}\| \quad \forall p = 2, 3, \cdots,$$

which by (4.5) entails

$$\|\omega\|_p \leqq |\Omega|^{1/p}\left(\frac{2}{\varepsilon}\|f\|_\infty + \frac{64\pi^2 c_4^2 c_0^{1/2}}{\varepsilon^3 \lambda_1^{1/2}} |f| K^{1/2}(\varepsilon, \lambda_1)\right) \quad \forall p = 2, 3, \cdots. \tag{4.6}$$

Passing to the limit, we deduce (4.3). The derivations of (4.2) and (4.4) follow simply from (3.3) and (3.14). We will omit their proofs.

In conclusion, the diagonal subsequence $\{\omega_{k_k}\}$ converges weakly to $\omega \in L^p(\Omega)$ for every $p = 2, 3, \cdots$, and $\{\psi_{k_k}\}$ converges weakly to $\psi \in H^2(\Omega)$ and strongly in $H^1(\Omega)$. Because $\nabla\psi_{k_k}$ converges strongly in $L^2(\Omega)$ to $\nabla\psi$ and ω_{k_k} converges weakly in $L^2(\Omega)$, it follows that the product $(\nabla\psi_{k_k})(\omega_{k_k})$ converges to $(\nabla\psi)(\omega)$ in the distribution sense. Since $\{\omega_{k_k}, \psi_{k_k}\}$ solves (3.1) for $\nu = \nu_k$, by passing to the limit we can verify that (ω, ψ) solves (4.1).

Up to now, the parameter R did not enter at all in our estimates. In the next theorem, we give an upper bound to the "diameter" of the set of stationary solutions of problem (4.1).

THEOREM 4.2. *Let* (ψ_1, ω_1) *and* (ψ_2, ω_2) *be two weak solutions to problem* (4.1) *satisfying* (4.2)–(4.4). *Then*

$$\|\psi_1 - \psi_2\| \leqq \frac{4R}{\varepsilon} |f|^2 K(\varepsilon, \lambda_1)\left(\frac{2}{\varepsilon}\|f\|_\infty + \frac{64\pi^2 c_4^2 c_0^{1/2}}{\varepsilon^3 \lambda_1^{1/2}} |f| K^{1/2}(\varepsilon, \lambda_1)\right). \tag{4.7}$$

Proof. From (4.1) we have

$$\varepsilon\Delta(\psi_1-\psi_2)+\frac{\partial}{\partial x_1}(\psi_1+\psi_2)+R\ \mathrm{div}\,(\omega_1\mathbf{u}_1-\omega_2\mathbf{u}_2)=0,$$

or

$$\varepsilon\Delta(\psi_1+\psi_2)+\frac{\partial}{\partial x_1}(\psi_1-\psi_2)+R\ \mathrm{div}\,(\omega_1(\mathbf{u}_1-\mathbf{u}_2)+(\omega_1-\omega_2)\mathbf{u}_2)=0.$$

We form the scalar product with $(\psi_1-\psi_2)$ and, since $\psi_1-\psi_2\in H_0^1(\Omega)$, we can integrate by parts and obtain

$$\varepsilon\|\psi_1-\psi_2\|^2+R(\mathbf{u}_2(\omega_1-\omega_2),\nabla(\psi_1-\psi_2))=0. \tag{4.8}$$

From (4.8) we get

$$\varepsilon\|\psi_1-\psi_2\|^2\leqq R\|\omega_1-\omega_2\|_\infty|\mathbf{u}_2|\,\|\psi_1-\psi_2\|,$$

and hence

$$\|\psi_1-\psi_2\|\leqq\frac{R}{\varepsilon}\|\omega_1-\omega_2\|_\infty\|\psi_2\|. \tag{4.9}$$

As a consequence of Lemma 3.1 and Theorem 4.1,

$$\|\psi_2\|\leqq 2|f|^2K(\varepsilon,\lambda_1).$$

Substituting this expression and (4.3) in (4.9) we arrive at (4.7).

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Elsevier, New York, 1965.

[3] F. E. BROWDER, *Fixed point theory and nonlinear problems*, Bull. Amer. Math. Soc. (N.S.), 9 (1983), pp. 1-39.

[4] K. BRYAN, *A numerical investigation of a non-linear model of wind-driven ocean*, J. Atmospheric Sci., 20 (1963), pp. 594-606.

[5] J. G. CHARNEY, *The Gulf Stream as an inertial boundary layer*, Proc. Nat. Acad. Sci. U.S.A., 41 (1955), pp. 731-740.

[6] P. CONSTANTIN AND C. FOIAS, *Navier-Stokes Equations*, The University of Chicago Press, Chicago, 1988.

[7] O. D. KELLOGG, *Foundations of Potential Theory*, Springer-Verlag, New York, 1967.

[8] J. L. LIONS AND E. MAGENES, *Problèmes aux limites nonhomogènes et applications*, Dunod, Paris, 1968-1970.

[9] G. MINEA, *Remarque sur les équations d'Euler dans un domaine possédant une symmetrie de revolution*, C.R. Acad. Sci. Paris, *Sér. I Math.*, 284 (1977), pp. 477-479.

[10] W. H. MUNK AND G. F. CARRIER, *The wind-driven circulation in ocean basins of various shapes*, Tellus, 2 (1950), pp. 158-167.

[11] J. PEDLOSKY, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1979.

[12] J. PEDLOSKY AND H. P. GREENSPAN, *A simple laboratory model for the ocean circulation*, J. Fluid Mech., 27 (1967), pp. 291-304.

[13] H. STOMMEL, *The westward intensification of wind-driven ocean currents*, Trans. Amer. Geophys. Union, 29 (1948), pp. 202-206.

[14] R. TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis*, CBMS-NSF Regional Conference in Applied Mathematics 41, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1983.

[15] G. VERONIS, *Wind-driven ocean circulation—Part 1. Linear theory and perturbation analysis*, Deep-Sea Res., 13 (1966), pp. 17-29.

[16] ———, *Wind-driven ocean circulation—Part 2. Numerical solution of the non-linear problem*, Deep-Sea Res., 13 (1966), pp. 31-55.

[17] V. I. YUDOVITCH, *Non-stationary flow of an ideal incompressible liquid*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1963), pp. 1032-1066.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

MULTIPLE TRAVELING WAVES IN A COMBUSTION MODEL*

S. P. HASTINGS†

Abstract. A model reaction scheme, consisting of two simple competing reactions $A \to P_1$ and $A \to P_2$, is studied using Arrhenius kinetics with a cut-off to handle the cold boundary difficulty. It is shown that for appropriate values of the parameters in the problem, the model equations have three distinct traveling wave solutions. The middle solution, presumably unstable, is obtained from a singularly perturbed problem by rigorous matching.

Key words. combustion, traveling waves

AMS(MOS) subject classifications. 35K57, 80A3D

1. Introduction. Recently there has been progress in the mathematical analysis of planar fronts in premixed combustion, thereby putting some rigorous buttressing around the walls of high activation energy asymptotics. For example, there are existence theorems for the simple reaction scheme $R \to P$, where R is the reactant and P the product, and also results justifying the inner and outer expansions developed by formal methods [1], [5], [8]. Similar results have been obtained for a chain reaction with two steps [7]. However, these models do not support strict traveling waves, unless artifical modifications are made in the usual Arrhenius reaction rate, because of the cold boundary difficulty [2]. By introducing intermediate variables, or radicals, this problem disappears, but the mathematical difficulties are considerably greater. Again, some progress has been made in proving existence of a flame [4], but much remains to be done.

One phenomenon that has not yet received a rigorous treatment is the possibility of several flames existing under the same conditions of temperature and concentration of the unburned mixture. Formal analysis was carried out by Clavin, Fife, and Nicolaenko [3], for the case of two competing reactions, with the scheme

(i) $R \to P_1$,
(ii) $R \to P_2$.

It was found that if the activation energies and rates of heat release of the two reactions are ordered correctly, then for a range of unburned reactant concentrations, two flames traveling at different speeds are possible. The relevant equations actually have three distinct solutions, but only two of these seem to be stable and therefore representative of real flames.

This is perhaps the simplest model from premixed combustion where such multiplicity is found, and our purpose here is to demonstrate rigorously the existence of at least three solutions. A defect in our result is that the simple model (i)-(ii) again suffers from the cold boundary difficulty, and the reaction rate term must be modified in order for any solution to exist. We hope eventually to combine the techniques used here with those from [4] to obtain a multiplicity result for a reaction scheme involving intermediate species and unmodified Arrhenius kinetics.

2.1. Statement of the result for simple kinetics. Let $k_1(T)$ and $k_2(T)$ define the temperature dependence of the reaction rates for (i) and (ii), respectively. It is assumed that there is a cutoff temperature T_-, below which k_1 and k_2 are both zero. Further,

* Received by the editors September 21, 1987; accepted for publication January 11, 1988. This research was supported by the National Science Foundation.

† Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

it is assumed that there is a crossover temperature T^* such that $k_1(T) < k_2(T)$ for $T_- < T < T^*$ and $k_1(T) > k_2(T)$ for $T > T^*$. This results in expressions of the form

$$k_i(T) = e^{\theta_i(T-T^*)/TT^*} \quad \text{for } i = 1, 2, \text{ and } T > T_-$$

where $\theta_1 > \theta_2 > 0$. (It is convenient, but not necessary, for us to assume that both reactions are cut off at the same temperature.)

By appropriate rescaling, the differential equations become

$$Ley'' = y' + Dyk_1(T) + Dyk_2(T), \tag{1}$$

$$T'' = T' - Q_1 Dyk_1(T) - Q_2 Dyk_2(T), \tag{2}$$

where Le, Q_1, and Q_2 are given positive constants and D is an adjustable parameter. The boundary conditions to be considered are

$$y(-\infty) = y_u, \qquad T(-\infty) = 0 \tag{3}$$

and

$$y(\infty) = 0, \qquad T'(\infty) = 0, \tag{4}$$

where $y_u > 0$ is given. It is easily shown that the conditions (4) imply that $\lim_{x\to\infty} T(x)$ exists. However, this number, denoted by T_b for "burned state," is not specified ahead of time.

By a "solution" of (1)-(4) we mean a triple (D, y, T) where y and T satisfy the equations for the given value of D.

THEOREM. *Suppose that $Q_1 y_u > T^* > Q_2 y_u$. For any μ with $0 < \mu < 1$, there is a θ^* such that if $\theta^* \leqq \theta_2 \leqq \mu\theta_1$, then there are at least three distinct solutions of* (1)-(4).

2.2. Preliminary calculations and outline of proof. Henceforth we shall always assume that μ has been chosen and that $\theta_2 \leqq \mu\theta_1$. We use a shooting method, starting from $x = 0$, where x is the independent variable. We assume always that $T(0) = T_-$, $T'(0) = T_-$. Therefore $T(x) = T_- e^x$ for $x \leqq 0$, and T automatically satisfies the correct condition at $-\infty$. Also, if $y(0) = y'(0) + y_u$, then $y(x) = y'(0)e^x + y_u$ for $x \leqq 0$, and y also satisfies the desired condition at $-\infty$.

LEMMA 1. *Let*

$$\Gamma = \{(D, y'(0)) \mid y \text{ decreases on } (-\infty, \infty) \text{ and } y(+\infty) = 0\}.$$

Then Γ contains a continuum γ, which, for any $\bar{D} > \underline{D} > 0$, connects the line $D = \bar{D}$ with the line $D = \underline{D}$.

The proof will be given in § 2.3.

Let Δ denote the point $(D, y'(0))$. We assume from now on that Δ lies in γ. The problem is to choose this point so that $T'(\infty) = 0$. For each $D > 0$, let

$$\gamma_D = \{(\tilde{D}, y'(0)) \in \gamma \mid \tilde{D} = D\}.$$

We now define two subsets of the continuum γ:

$$A = \{\Delta \mid \text{for some } x > 0,\ T'(x) < 0\}$$

and

$$B = \{\Delta \mid \text{for some } x > 0,\ T(x) > T_1 = Q_1 y_u\}.$$

The following results will also be proved below.

LEMMA 2. *If D is positive but Δ does not lie in $A \cup B$, then the corresponding solution satisfies* (1)-(4).

LEMMA 3. *The sets A and B are open and disjoint.*

LEMMA 4. *If D is sufficiently large (depending on θ_1 and θ_2), then $\gamma_D \subset A$ and T' becomes negative before $T = T_2 \equiv Q_2 y_u$, while if D is positive but sufficiently small, then $\gamma_D \subset B$. Furthermore, there is an $\eta > 0$ such that for sufficiently large θ_2 and $\theta_2 \leqq \mu\theta_1$, T' cannot have a zero at a point where $T^* - \eta < T < T^* - \eta/2$.*

LEMMA 5. *For each sufficiently large θ_1, the set A is a disconnected subset of γ with at least three boundary points in γ.*

Lemma 5 is by far the most difficult of these results to prove. Lemmas 2-5 imply that there are at least three points in γ which do not lie in either A or B, and each of these corresponds to a solution of (1)-(4). Hence they imply the truth of our theorem.

2.3. Proofs of Lemmas 1–4. The first assertion of Lemma 1 follows from a straightforward shooting argument. We observe that y and y' cannot vanish simultaneously, unless y is identically zero. Also, if $y' = 0$ and $T > T_-$, then sign $(y'') =$ sign (y). This implies that the planar set

$$\Omega = \{(D, y'(0)) | D > 0,\ y'(0) > -y_u,$$
$$\text{and } y'(x) > 0 \text{ for some positive } x\}$$

is open and disjoint from the open set

$$\Lambda = \{(D, y'(0)) | D > 0,\ y'(0) > -y_u,$$
$$\text{and } y(x) < 0 \text{ for some } x > 0\}.$$

Furthermore, for fixed $D > 0$, if $y'(0) > 0$ then $(D, y'(0))$ lies in Ω, while if $y'(0)$ is sufficiently close to $-y_u$, then this point lies in Λ. Lemma 1 now follows from a result in topology.

PROPOSITION [6]. *If S is a square and P and Q are disjoint open sets containing, respectively, the right and left sides of S, then there is a continuum connecting the top and bottom of S within S and this continuum lies in the complement of $A \cup B$.*

Lemmas 2 and 3 are easily proved, and we omit the details.

The proof of Lemma 4 is a little harder, and requires the use of some techniques to be employed for Lemma 5 as well. Integrate (1) and (2) from $-\infty$ to x, using the boundary conditions (3), which we know are satisfied by our choice of $T(0)$, $T'(0)$, and $y(0)$ as a function of $y'(0)$. We obtain

$$Ley'(x) = y(x) - y_u + \alpha(x) + \beta(x), \tag{5}$$

$$T'(x) = T(x) - Q_1\alpha(x) - Q_2\beta(x) \tag{6}$$

where

$$\alpha(x) = \int_{-\infty}^{x} Dy(s) k_1(T(s))\, ds$$

and

$$\beta(x) = \int_{-\infty}^{x} Dy(s) k_2(T(s))\, ds.$$

It is obvious from (6) that $T'(x) \leqq T_2$ as long as $T \leqq T_2$, so we can find an $x_1 > 0$ such that $T(x_1) \leqq (T_2 + T_-)/2$ for any $\Delta \in \gamma$. (Recall that $T(0) = T'(0) = T_- < T_2$.) When we let $D \to \infty$, with Δ remaining in γ, there are two possibilities: either $y(x_1) \to 0$, or there is a $\delta > 0$ and a sequence of D's tending to ∞ such that $y(x_1) \geqq \delta$. In the latter

case it is clear that $Dyk_1(T)$ tends to infinity uniformly on $[0, x_1]$, which implies that T' becomes negative. In the former case, $y(x_1)$ and $y'(x_1)$ must both tend to zero. From (5) we see that $\alpha(x_1)+\beta(x_1)\to y_u$. Then, from (6),

$$Q_1\alpha(x_1)+Q_2\beta(x_1)>Q_2(y_u-\varepsilon)>T(x_1),$$

for some small $\varepsilon>0$ and sufficiently large D, which again implies that $T'(x_1)<0$.

The assertion that if D is small then $\gamma_D\subset B$ is easier and we omit the proof. For the last statement, suppose that $T'(x_0)=0$. From (6), $Q_1\alpha(x_0)+Q_2\beta(x_0)=T(x_0)$ and $\alpha(x_0)+\beta(x_0)<y_u$. Since T is increasing up to x_0, $\beta(x_0)/\alpha(x_0)\geqq k_2(T(x_0))/k_1(T(x_0))$. The latter ratio tends to infinity with θ_2 if $\theta_2\leqq\mu\theta_1$ and $T(x_0)<T^*-\delta$, for any fixed δ. Combining these shows that for some small η and for large θ_2, $T(x_0)\leqq T^*-\eta$.

2.4. Proof of Lemma 5 for a special case. We begin by discussing a restricted set of parameters in order to concentrate on the main idea of the proof. Later the extension to a full range as described in the theorem will be covered. Our assumption for the moment is that

$$T_2<T^*<(T_1+T_2)/2. \tag{7}$$

There are several steps in this proof, as given in the following additional lemmas.

LEMMA 6. *Fix $D=D^*$, an arbitrary positive number. Then for sufficiently large θ_1 and θ_2, chosen independently of Δ in γ_{D^*}, and $\theta_2\leqq\mu\theta_1$, Δ lies in the set A, but T increases to above T^* before T' becomes negative.*

Proof. This is the key result. Use is made of (5) and (6). Let

$$Q(x)=Q_1\alpha(x)+Q_2\beta(x).$$

Roughly, it is observed from (6) that T' becomes negative when $Q_1\alpha+Q_2\beta$ exceeds T. From (5) it is seen that $\alpha(x)+\beta(x)\to y_u$, as $x\to\infty$. As long as $T<T^*$, β/α is large, so Q cannot grow much above $Q_2y_u=T_2$. The type of solution described in Lemma 5 is only possible if T increases beyond T^* in such a way that α/β becomes large enough for Q to exceed T.

Reference must be made to a certain limiting solution of (1)-(3). Let y_0 and T_0 be the unique solution to (1)-(3) when $D=0$ such that $y_0(x)=0$ when $T_0(x)=T^*$. Thus,

$$y_0(x)=-y_ue^{(x-x^*)/Le}+y_u$$

and

$$T_0(x)=T_-e^x,$$

where $x^*=\log(T^*/T_-)$.

LEMMA 7. *For any fixed positive D, $T\to T_0$ uniformly in $[0, x^*]$ and uniformly in γ as long as D is held constant, as $\theta_2\to\infty$.*

Proof. It is easy to see that $T(x)<T_0(x)$ and $T'(x)<T_0'(x)$ for all $x>0$. If the result is false, then there must be a $\delta>0$ such that for a sequence of θ_2's tending to infinity, $T(x)\leqq T^*-\delta$ on $[0, x^*]$. Then, however, $k_1(T(x))$ and $k_2(T(x))\to 0$ uniformly on this interval. The initial conditions of T are the same as those for T_0, irrespective of θ_1 and θ_2, and Δ. It follows that T must tend to T_0 as claimed.

COROLLARY 1. *$\alpha(x)\to 0$ and $\beta(x)\to 0$ as $\theta_2\to\infty$, uniformly on $[-\infty, x^*]$ and in γ.*

Proof. The functions α and β are nondecreasing, with derivatives bounded by Dy_u as long as $T\leqq T^*$. The result follows from (6) and Lemma 7, since $T_0'=T_0$.

In the following result, D is still held constant.

LEMMA 8. *As $\theta_2\to\infty$, $y(x)\to y_0(x)$ uniformly on $[0, x^*]$.*

Proof. We first show that $y'(0)\to y_0'(0)$. First suppose that there is a $\delta>0$ such that, for arbitrarily large values of θ_2, $y'(0)\leqq y_0'(0)-\delta$. From Lemma 7 it follows that

the $k_i(T(x))$ tend to zero uniformly on compact subintervals of $[0, x^*)$, and this implies that y becomes negative for sufficiently large θ_2, because the solution of $y''=y'$, $y'(0)=y_0'(0)-\delta$, $y(0)=y'(0)+y_u$ becomes negative on the interval $[0, x^*]$. This contradicts the way $y'(0)$ is chosen.

Next, we suppose that for some $\delta>0$,

$$y'(0)\geqq y_0'(0)+\delta,$$

for an unbounded sequence of (positive) θ_2's. Then for these θ_2's,

$$y(x^*)\geqq\delta. \tag{8}$$

Also, $|y'|\leqq y_u$. Hence we see that

$$y(x)\geqq\delta-y_u(x-x^*) \tag{9}$$

on $I=[x^*, x^*+\delta_1]$, where $\delta_1=\delta/y_u$.

LEMMA 9. *Under our present assumptions, at least one of the terms $Dy(x)k_i(x)$ is unbounded on the interval $I_2=[x^*, x^*+\delta_1/2]$, that is, $\max_{x\in I_2}|Dyk_i(x)|\to\infty$ as $\theta\to\infty$.*

Proof. Suppose not, and there is an M such that $|Dy(x)k_i(x)|\leqq M$ on I_2, for each i. Then on this interval $|Q(x)|\leqq o(1)+M_1(x-x^*)$, where $M_1=m(Q_1+Q_2)$ and where $o(1)$ is a term which tends to 0 as $\theta_2\to\infty$. (We use the corollary above.) From (6) it then follows that unless T' becomes negative there is an $\varepsilon>0$ such that for large θ_2, $T(x)\geqq T^*+\varepsilon$ on $I_3=[x^*+\delta_1/4, x^*+\delta_1/2]$. However, this implies that $k_1(T)\to\infty$ uniformly on this interval, and this with (10) contradicts the assumption that $|Dyk_i|\leq M$ on I_2.

COROLLARY 2. *For at least one i, $Dyk_i(T)\to\infty$ as $\theta\to\infty$, uniformly in $[x^*+\delta_1/2, x^*+\delta_1]$.*

Proof. This follows because $d(yk_i(T))/dx$ is bounded below.

This corollary, with (9), leads immediately to the conclusion that y' becomes positive. This is a contradiction, and so we have proved that $y'(0)\to y_0'(0)$, which is Lemma 8.

In fact, what we have shown is that (8) is impossible for large enough θ_2. This proves the following lemma.

LEMMA 10. *As $\theta_2\to\infty$, $y(x)\to y_0(x)$ uniformly on $[0, x^*]$, and $y'(x)\to y_0'(x)$ uniformly on any compact subinterval of $[0, x^*)$, for a fixed value of D, and $\Delta\in\gamma_D$.*

Since $y(x^*)=0$, $y'(x)<0$, and $y(x)>0$ for all x, we have demonstrated that there is a "corner layer" at x for large θ_2 and fixed D. The following limits are all uniform in γ_{D^*}.

LEMMA 11. $\lim_{\rho\to0}\lim_{\theta\to\infty}y'(x^*-\rho)-y'(x^*+\rho)=-y_u$.

Proof. This follows from Lemma 8, because $y_0'(x^*)=-y_u$.

COROLLARY 3. $\lim_{\rho\to0}\lim_{\theta\to\infty}\{(\alpha+\beta)|_{x^*+\rho}-(\alpha+\beta)|_{x^*-\rho}\}=y_u$.

Proof. This follows by integrating (5) from $x^*-\rho$ to $x^*+\rho$ and using the definitions of α and β.

LEMMA 12. $\lim_{\rho\to0}\liminf_{\theta\to\infty}\alpha(x^*+\rho)/\beta(x^*+\rho)\geqq1$.

Proof. This crucial result requires several steps to prove.

(i) $T(x^*)\to T^*$ as $\theta_2\to\infty$. If not, then there is a $\delta>0$ and sequence of θ_2's tending to infinity such that $T(x^*)\leqq T^*-\delta$. Then $T(x)\leqq T^*-\delta$ on $[0, x^*]$. But this implies that each $k_i(T(x))\to0$ uniformly on $[0, x^*]$, which in turn implies that $T(x^*)$ does tend to $T_0(x^*)$, a contradiction.

(ii) As $\theta_2\to\infty$, $\alpha(x^*)\to0$ and $\beta(x^*)\to0$. Choose some $\varepsilon>0$. Then there is a $\delta>0$ such that $T\leqq T^*-\delta$ on $[0, x^*-\varepsilon]$, for all large θ_2. Then $\alpha(x^*-\varepsilon)\to0$. Also, $\alpha(x^*)\leqq\alpha(x^*-\varepsilon)+\varepsilon Dy_u$ because $T(x^*)\leqq T^*$. The proof for β is the same.

(iii) There is an $\hat{x} > x^*$ such that $T(\hat{x}) = T^*$, and $\hat{x} \to x^*$ as $\theta_2 \to \infty$. For each $\delta > 0$, $T'(x^* - \delta) \to T_0'(x^* - \delta)$. Also, $T_0'(x) \geqq T_-$ and $T'' \geqq T - Q_1 D - Q_2 D$ as long as $T \leqq T^*$. Therefore T crosses T^* near x^*.

(iv) As $\theta_2 \to \infty$, $\alpha(\hat{x}) \to 0$ and $\beta(\hat{x}) \to 0$. This is shown in the same way as (ii). Lemma 12 now follows because $\theta_1 \geqq \theta_2$ and $T(x^* + \rho) > T^*$ for large θ_2 and any given ρ.

LEMMA 13. $\lim_{\rho \to 0} \lim_{\theta_2 \to \infty} Q(x^* + \rho) > T^*$.

This follows from Lemmas 8 and 12, Corollary 3, and inequality (7).

In fact, for large θ_2, both α and β remain about constant beyond x^*. More precisely, we have Lemma 14.

LEMMA 14. *For any* $\rho > 0$, $\lim_{\theta_2 \to \infty} (\alpha(x) - \alpha(x^* + \rho)) = 0$ *and* $\lim_{\theta_2 \to \infty} (\beta(x) - \beta(x^* + \rho)) = 0$, *uniformly in* $[x^* + \rho, \infty)$. *Furthermore,* $\alpha(x^* + \rho) + \beta(x^* + \rho) \to y_u$.

Proof. The second assertion was already proved. The first assertion follows because α and β are nondecreasing, and from (5), because $y(\infty) = 0$.

COROLLARY 4. $\lim_{\theta_2 \to \infty} (Q(x) - Q(x^* + \rho)) = 0$ uniformly on $[x^* + \rho, \infty)$, and $\lim_{\theta_2 \to \infty} Q(x^* + \rho)$ exists and is greater than T^*.

The final step in the proof of Lemma 6 for the special case under consideration is to solve the differential equation (6) starting at $x = x^*$. We obtain

$$T(x) = e^{x - x^*}[T^* - Q(x^* + \rho) + O(\rho) + o(1) + O(e^{-x})]$$

for small ρ and large x as $\theta_2 \to \infty$. The desired conclusion that T' becomes negative for sufficiently large θ_2 follows.

The remainder of the proof of Lemma 5 is relatively routine, and we merely outline the steps. Fix some small $\delta < T^* - T_2$. For $T \leqq T^* - \delta$, k_2/k_1 is small for large θ_2, if $\theta_2 \leqq \mu\theta_1$. Therefore the second reaction (ii) dominates. It is not hard to show, using simple shooting, that this reaction supports a traveling wave, with final temperature $T_b = T_2$. For large θ_2, we can show that the pair of reactions supports a flame with T_b close to T_2. More precisely, we have seen that if D is sufficiently large, then $\Delta \in A$ and $T' = 0$ before $T = T_1$. In particular, consider some unbounded component γ_1 of A. By Lemmas 2 and 4, γ_1 is entirely contained in the region $D > D^*$, since the point where $T' = 0$ depends continuously on initial conditions and cannot jump from below $T^* - \eta$ to above T^*. Some point of γ_{D^*} defines a second component γ_2 of A, separated from γ_1 in γ. If D is sufficiently small, then $\Delta \in B$. Therefore γ_2 has at least two boundary points, one with $D > D^*$ and one with D below D^*. Any boundary point corresponds to a solution of (1)-(4). Suppose that there are only two boundary points of γ_2, and let p be the one with $D > D^*$.

LEMMA 15. *The point p is not a limit point of the set* γ_1.

Proof. p is a limit point of points in γ_2, and all these points correspond to solutions where T has a maximum above T^*. Therefore p corresponds to a solution such that $T(\infty) \geqq T^*$. It follows that nearby solutions must also reach at least $T^* - \eta/2$. This is impossible for points in γ_1.

It follows, therefore, that γ_1 has a boundary point different from any boundary point of γ_2. This proves Lemma 5 and our theorem in the special case.

2.5. The general case. We wish to remove the restriction (7), and allow any T^* between T_1 and $T_2 = Q_2 y_u$. This is accomplished by improving the estimate on α/β. In order to achieve a sufficiently large ratio α/β, it is clearly necessary that T increases beyond T^*. The statement to this effect in Lemma 6 must now be more precise.

Lemma 16. *Let* $K = ((T_1 - T_2)/(T_1 - T^*)) - 1$. *Choose a number* λ *so that*

$$e^{\lambda(1-\mu)} > K.$$

Then for sufficiently large θ_2, $(T - T^*)/TT^*$ *must exceed* λ/θ_1, *at some* x^{**} *such that* $x^{**} \to x^*$ *as* $\theta_2 \to \infty$. *Furthermore,* $\alpha(x^{**}) \to 0$ *and* $\beta(x^{**}) \to 0$ *as* $\theta_2 \to \infty$.

Proof. This improved estimate is obtained by modifying the argument for Corollary 1 of Lemma 7. The derivatives of α and β are bounded independent of θ_1, for fixed D, as long as $(T - T^*)/TT^* \leqq \lambda/\theta_1$. (Here λ is fixed as above, while $\theta_1 \to \infty$.) Therefore the same reasoning as used in Corollary 1 yields this improved result.

Now all the arguments subsequent to Corollary 1 are the same, if x^* is replaced by x^{**}. Note that

$$k_1/k_2 \geqq e^{\lambda(1-\mu)}$$

for $x \geqq x^{**}$. Therefore, for given $\rho > 0$ and sufficiently large θ, $\alpha(x^{**} + \rho)/\beta(x^{**} + \rho) \geqq K$. The definition of K, and the further result that $(\alpha + \beta)|_{x^{**}+\rho} \to y_u$ as $\theta \to \infty$, imply that $Q(x^{**} + \rho) > T^*$ for sufficiently large θ_2, and the result follows as before.

REFERENCES

[1] H. BERYSTICKII, B. NICOLAENKO, AND B. SCHEURER, *Traveling wave solutions to reaction-diffusion systems modelling combustion*, SIAM J. Math. Anal., 16 (1985), pp. 1207–1242.

[2] J. BUCKMASTER AND G. S. S. LUDFORD, *Theory of Laminar Flames*, Cambridge University Press, London, 1982.

[3] P. CLAVIN, P. FIFE, AND B. NICOLAENKO, *Multiplicity and related phenomena in competing reaction flames*, SIAM J. Appl. Math., 47 (1987), pp. 296–331.

[4] S. HASTINGS, C. LU, AND Y.-H. WAN, *Existence of traveling waves in a model with no cold boundary difficulty*, SIAM J. Appl. Math., 47 (1987), 1229–1240.

[5] M. MARION, *Mathematical study of a model with no ignition temperature for laminar plane flames*, in Reacting Flows: Combustion and Chemical Reactors II, Lectures in Applied Mathematics 24, American Mathematical Society, Providence, RI, 1986.

[6] J. B. MCLEOD AND J. SERRIN, *The existence of similar solutions for some boundary layer problems*, Arch. Rational Mech. Anal., 31 (1968), pp. 268–303.

[7] D. TERMAN, *An application of Conley index to combustion*, NATO ASI Series, Vol. F37, Dynamics of Infinite Dimensional Systems, F.-N. Chow and J. K. Hale, eds., Springer-Verlag, Berlin, New York, 1987.

[8] D. WAGNER, *Premixed laminar flames as travelling waves*, in Reacting Flows: Combustion and Chemical Reactors II, Lectures in Applied Mathematics 24, American Mathematical Society, Providence, RI, 1986.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

010

MAXWELL EQUATIONS IN POLARIZABLE MEDIA*

BERNARDO COCKBURN† AND PATRICK JOLY†

Abstract. The resolution of Maxwell equations in polarizable conductive media led to the resolution of a linear integrodifferential system. A method for the numerical approximation of this system was proposed in [B. Cockburn, *SIAM J. Sci. Statist. Comput.*, 6 (1985), pp. 843-852]. Here the mathematical results that justify this method are given.

Key words. Maxwell equations, polarizable media, integrodifferential systems, nonlocal operators, approximation by rational fractions

AMS(MOS) subject classifications. 41A20, 45K05, 78A99

Introduction. In [2], Cockburn proposed and described a method for the numerical approximation for the solution of the following integrodifferential system:

$$\mu \frac{\partial u}{\partial t}(x, t)+\frac{\partial v}{\partial x}(x, t)=0, \tag{0.1}_1$$

$$\sigma_\infty(x)v(x, t)+\int_0^t \tilde{\sigma}_1(x, t-s)v(x, s)\, ds+\frac{\partial u}{\partial x}(x, t)=0. \tag{0.1}_2$$

From both mathematical and numerical points of view, the main difficulty lies in the treatment of the nonlocal time-convolution operator appearing in $(0.1)_2$. Such a system arises when we consider the propagation of an electromagnetic field $(\mathbf{E}, \mathbf{H})$ in a one-dimensional conductive and polarizable medium. $(\mathbf{E}, \mathbf{H})$ satisfies the Maxwell equations that can be written in the space-frequency $(\mathbf{x}=(x, y, z), \omega)$ domain as follows:

$$\operatorname{curl} \mathbb{E}(\mathbf{x}, \omega)=i\mu\omega\, \mathbb{H}(\mathbf{x}, \omega), \qquad \operatorname{curl} \mathbb{H}(\mathbf{x}, \omega)=\sigma(x, \omega)\mathbb{E}(\mathbf{x}, \omega)$$

(here $\mathbb{E}$ and $\mathbb{H}$ denote the time Fourier transforms of $\mathbf{E}$ and $\mathbf{H}$).

The function $\sigma(x, \omega)$ is the (complex) conductivity of the medium. The dependence of σ on the frequency ω is determined by the polarizability of the medium (see Diaz [5] or Goldman [6] for further details concerning the physical model).

One of the most classical polarization laws is Warbourg's law, which corresponds to the formula

$$\sigma(\omega)=\sigma_0\left(\left(1+\lambda\left(i\frac{\omega}{\omega_c}\right)^{1/2}\right)\Big/\left(1+\left(i\frac{\omega}{\omega_c}\right)^{1/2}\right)\right) \qquad (\operatorname{Re}(\mathbf{Z}^{1/2})>0). \tag{0.2}$$

If the data are independent on y and z it can be shown that the fields $\mathbf{E}$ and $\mathbf{H}$ can be expressed as follows:

$$\mathbf{E}(\mathbf{x}, t)=E(x, t)\vec{y}_0, \qquad \mathbf{H}(\mathbf{x}, t)=H(x, t)\vec{z}_0.$$

Assuming that

$$\sigma(x, \omega)=\sigma_\infty(x)+\frac{1}{\sqrt{2\pi}}\int_0^{+\infty} \tilde{\sigma}_1(x, t)\, e^{-i\omega t}\, dt,$$

we then easily show that the pair $(u, v)=(E, H)$ is a solution of the system (0.1).

Of course, for the two- and three-dimensional cases, the fields $\mathbf{E}$ and $\mathbf{H}$ are solutions of an integrodifferential system analogous to (0.1) and our analysis will still be valid.

* Received by the editors March 3, 1986; accepted for publication February 3, 1988.

† Institut National de Recherche en Informatique en Automatique, Domaine de Voluceau-Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France.

For the sake of simplicity, in this article we will only present results concerning the system (0.1).

The idea developed in [2] for treating the convolution operator consists in approximating this nonlocal operator by a local one. This can be achieved by approximating the kernel $t \to \tilde{\sigma}_1(x, t)$ by a sum of decreasing exponential functions, or equivalently, by approximating the time Fourier transform $\sigma_1(x, \omega)$ of this kernel by a sum of simple rational fractions.

More precisely, if we approximate $\sigma(x, \omega)$ by the following function:

$$\sigma_\infty(x) + \sum_{k=1}^{n} \frac{b_k(x)}{1 + i a_k(x)\omega}, \qquad b_k(x) > 0, \quad a_k(x) > 0, \quad k = 1, 2, \cdots, n, \tag{0.3}$$

then, as will be proved in §2, the system (0.1) can be approximated by the new system

$$\begin{cases} \mu \dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, \\ \sigma_\infty v + \displaystyle\sum_{k=1}^{n} f_k + \frac{\partial u}{\partial x} = 0, \\ a_k \dfrac{\partial f_k}{\partial t} + f_k = b_k u, \qquad k = 1, 2, \cdots, n, \end{cases} \tag{0.4}$$

where the functions f_k are auxiliary functions related to the approximate kernel (0.3).

Thus, we have replaced a system of two integrodifferential equations by a system of $(n+2)$ differential equations, the numerical approximation of which is now classical (see [2], [3]).

In this article, our aim is to give some theoretical justifications of this approach. In § 1 we establish some preliminary results about the time convolution operator. In § 2 we give the mathematical analysis of the system (0.1) (existence, uniqueness, regularity, and continuity results with respect to σ).

Finally, in § 3, we provide a complete justification of our method in the case where the polarization is given by Warbourg's law (0.2). More precisely, we give the following: (1) A constructive process for obtaining the approximate polarization's law; and (2) An error estimate for the approximation of the solution of (0.1) by the solution of (0.4).

1. Mathematical properties of the convolution operator. Let us introduce the following notation. For $I \subset \mathbb{R}$, $\mathbb{L}^p(I)$ $(1 \leqq p < +\infty)$ will denote the space of L^p functions with complex values and $L^p(I)$ the subspace of $\mathbb{L}^p(I)$ of functions with real values. We will identify $\mathbb{L}^p(\mathbb{R}^+)$ with the subspace of functions $f(t)$ in $\mathbb{L}^p(\mathbb{R})$ that are 0 equal for $t < 0$. By $\mathscr{F}$ we will denote the Fourier transform in $\mathbb{L}^2(\mathbb{R})$ defined for regular functions $u(t)$ by

$$(\mathscr{F}u)(\omega) = \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{+\infty} u(t)\, e^{-i\omega t}\, dt.$$

($\mathscr{F}$ is an isometry in $\mathbb{L}^2(\mathbb{R})$.)

As a first step, we assume that the convolution kernel σ does not depend on the space variable x and we study the convolution operator formally defined by

$$(\phi_\sigma u)(t) = \sigma_\infty u(t) + \int_0^t \sigma_1(t-s) u(s)\, ds. \tag{1.1}$$

1.1. Study of the operator ϕ_σ.

1.1.1. The class of admissible polarization laws. We introduce the following set of complex functions:

$$\Sigma_\omega(\mathbb{R}) = \{\sigma(\omega) : \mathbb{R} \to \mathbb{C} \text{ satisfying (1.3), (1.4)}\}, \tag{1.2}$$

$$\exists(\sigma_\infty, \tilde{\sigma}_1(t)) \in \mathbb{R}_*^+ \times L^1(\mathbb{R}^+) \quad \forall \omega \in \mathbb{R} \quad \sigma(\omega) = \sigma_\infty + \frac{1}{\sqrt{2\Pi}} \int_{-\infty}^{+\infty} \tilde{\sigma}_1(t)\, e^{-i\omega t}\, dt, \tag{1.3}$$

$$\exists \sigma_* \in \mathbb{R}_*^+ \quad \forall \omega \in \mathbb{R} \quad \mathcal{R}e(\sigma(\omega)) \geqq \sigma_* > 0. \tag{1.4}$$

It is easy to verify that $\Sigma_\omega(\mathbb{R})$ has the following properties:

(1) $\Sigma_\omega(\mathbb{R})$ is an open, convex subset of the space $C_b^0(\mathbb{R})$ of complex functions that are continuous and bounded (with the L^∞ norm).

(2) For any σ in $\Sigma_\omega(\mathbb{R})$, we have the following:

$$\|\sigma\|_{L^\infty(\mathbb{R})} \leqq \sigma^* \qquad (\sigma^* = \sigma_\infty + \|\tilde{\sigma}_1\|_{L^1(\mathbb{R}^1)}),$$

$$\lim_{\omega \to \infty} \sigma(\omega) = \sigma_\infty,$$

$$\sigma(-\omega) = \overline{\sigma(\omega)},$$

$$|\sigma(\omega)| \geqq \mathcal{R}e(\sigma(\omega)) \geqq \sigma_*.$$

1.1.2. The operator $\hat{\phi}_\sigma$. We introduce the linear operator for $\sigma \in \Sigma_\omega(\mathbb{R})$:

$$\hat{\phi}_\sigma : \mathbb{L}^2(\mathbb{R}) \to \mathbb{L}^2(\mathbb{R}),$$

$$U(\omega) \to V(\omega) = (\hat{\phi}_\sigma U)(\omega),$$

$$V(\omega) = \sigma(\omega) U(\omega) \quad \text{a.e. } \omega \in \mathbb{R}.$$

Using the properties of σ it is easy to obtain the following Theorem.

THEOREM 1.1. *$\hat{\phi}_\sigma$ is a linear continuous operator in $\mathbb{L}^2(\mathbb{R})$, with continuous inverse $\hat{\phi}_\sigma^{-1}$. Moreover, we have*

$$\forall U \in \mathbb{L}^2(\mathbb{R}) \quad \sigma_* \|U\|_{L^2}^2 \leqq \mathcal{R}e(\hat{\phi}_\sigma U, U)_{L^2} \leqq \sigma^* \|U\|_{L^2}^2.$$

COROLLARY 1.1. *We have the following estimates*:

$$\|\hat{\phi}_\sigma\| \leqq \sigma^*, \qquad \|\hat{\phi}_\sigma^{-1}\| \leqq (\sigma_*)^{-1}.$$

1.1.3. The operator ϕ_σ in $\mathbb{L}^2(\mathbb{R})$. For every σ in $\Sigma_\omega(\mathbb{R})$, we define $\phi_\sigma = \mathcal{F}^{-1} \circ \hat{\phi}_\sigma \circ \mathcal{F}$

$$\begin{array}{ccc} \mathbb{L}^2(\mathbb{R}) & \xrightarrow{\phi_\sigma} & \mathbb{L}^2(\mathbb{R}) \\ \mathcal{F} \downarrow & & \downarrow \mathcal{F} \\ \mathbb{L}^2(\mathbb{R}) & \xrightarrow{\hat{\phi}_\sigma} & \mathbb{L}^2(\mathbb{R}) \end{array}$$

Since the operator $\mathcal{F}$ is unitary, we clearly have the following theorem.

THEOREM 1.2. *ϕ_σ is a continuous linear mapping in $\mathbb{L}^2(\mathbb{R})$. ϕ_σ is invertible and ϕ_σ^{-1} is continuous. Moreover,*

$$\forall u \in \mathbb{L}^2(\mathbb{R}) \quad \sigma_* \|u\|_{L^2}^2 \leqq \mathcal{R}e(\phi_\sigma u, u)_{L^2} \leqq \sigma^* \|u\|_{L^2}^2,$$

$$\|\phi_\sigma\| \leqq \sigma^* \quad \textit{and} \quad \|\phi_\sigma^{-1}\| \leqq (\sigma_*)^{-1}.$$

Using the properties of the Fourier transform with respect to the convolution and the fact that $\tilde{\sigma}_1$ is a causal integrable function, we easily establish the following theorem.

THEOREM 1.3. *For any u in $\mathbb{L}^2(\mathbb{R})$, $v = \phi_\sigma u$ is given by*

$$v(t) = \sigma_\infty u(t) + \int_{-\infty}^{t} \tilde{\sigma}_1(t-s)u(s)\,ds.$$

Knowing that $\tilde{\sigma}_1$ is a real function we deduce the following corollary.

COROLLARY 1.2. *ϕ_σ is a bijective operator from $L^2(\mathbb{R}^2)$ to itself. It maps $\mathbb{L}^2(\mathbb{R}^+)$ to $\mathbb{L}^2(\mathbb{R}^+)$ and when u belongs to $\mathbb{L}^2(\mathbb{R}^+)$*

$$\forall u \in \mathbb{L}^2(\mathbb{R}^+) \quad (\phi_\sigma u)(t) = \sigma_\infty(t) + \int_0^t \tilde{\sigma}_1(t-s)u(s)\,ds.$$

In this last formula, we can see that when the function u is causal, the value of $\phi_\sigma u$ at time t only depends on the values of u in the interval $[0, t]$ (we then say that the operator ϕ_σ is causal). This allows us to define the operator ϕ_σ on $L^2(0, T)$.

1.1.4. The operator ϕ_σ on $L^2(0, T)$. We now define

$$\phi_\sigma : L^2(0, T) \to L^2(0, T),$$
$$u \to \phi_\sigma u,$$
$$\phi_\sigma u(t) = \sigma_\infty u(t) + \int_0^t \tilde{\sigma}_1(t-s)u(s)\,ds.$$

The main result of this section is the following theorem.

THEOREM 1.4. *ϕ_σ is a linear, continuous, bijective mapping from $L^2(0, T)$ to $L^2(0, T)$, with continuous inverse ϕ_σ^{-1}. Moreover*

$$\forall u \in L^2(0, T) \quad \sigma_* \int_0^T |u(t)|^2 dt \leqq \int_0^T \phi_\sigma u(t)u(t)\,dt \leqq \sigma^* \int_0^T |u(t)|^2\,dt, \tag{1.5}$$

$$\|\phi_\sigma\| \leqq \sigma^* \quad \text{and} \quad \|\phi_\sigma^{-1}\| \leqq (\sigma_*)^{-1}$$

Proof of (1.5). Let u be in $L^2(0, T)$. We define the function $u^*(t)$ in $L^2(\mathbb{R})$ by

$$\begin{aligned} &u^*(t) = u(t) \quad \text{if } t \in [0, T], \\ &u^*(t) = 0 \qquad \text{if } t \notin [0, T]. \end{aligned} \tag{1.6}$$

Now we set

$$v = \phi_\sigma u \in L^2(0, T), \qquad v^* = \phi_0 u^* \in L^2(\mathbb{R}).$$

Since ϕ_σ is causal, we know that

$$\text{a.e.} \quad t \in [0, T] \quad v^*(t) = v(t). \tag{1.7}$$

From Theorem 1.2, we have

$$\sigma_*\left(\int_{-\infty}^{+\infty} |u^*(t)|^2\,dt\right) \leqq \int_{-\infty}^{+\infty} u^*(t)v^*(t)\,dt \leqq \sigma^*\left(\int_{-\infty}^{+\infty} |u^*(t)|^2\,dt\right).$$

But, thanks to (1.6) and (1.7), we have

$$\int_{-\infty}^{+\infty} u^*(t)v^*(t)\,dt = \int_0^T u(t)v(t)\,dt,$$
$$\int_{-\infty}^{+\infty} |u^*(t)|^2\,dt = \int_0^T |u(t)|^2\,dt.$$

Then (1.5) follows immediately. □

Remarks. It is clear that the inequalities (1.5) still hold if we replace T by any t in the interval $[0, t]$:

$$\forall t \in [0, T] \quad \sigma_* \int_0^t \|u(s)\|^2 \, ds \leqq \int_0^t \phi_\sigma u(s) u(s) \, ds \leqq \sigma^* \int_0^t \|u(s)\|^2 \, ds.$$

Note that the essential assumption to obtain the properties of ellipticity for ϕ_σ is

$$\mathcal{R}e(\sigma(\omega)) \geqq \sigma_* > 0 \quad \forall \omega > 0.$$

Finally, let us give a property of the operator ϕ_σ with respect to time derivation.

THEOREM 1.5. *For any function u in $H^1(0, T)$, we have*

$$\frac{d}{dt}(\phi_\sigma u) = \phi_\sigma\left(\frac{du}{dt}\right) + u(0)\tilde{\sigma}_1 \quad \text{in } \mathcal{D}'(0, t).$$

This identity is easily obtained in the sense of distributions. Note that this equality makes sense since $H^1(0, T) \hookrightarrow C^0(0, T)$ and $\tilde{\sigma}_1 \in L^1(0, T)$, so that $u(0)\,\tilde{\sigma}_1$ belongs to $L^1(0, T)$ and thus to $\mathcal{D}'(0, T)$.

1.1.5. Continuity of the mapping $\sigma \to \phi_\sigma$. $\Sigma_\omega(\mathbb{R})$ is a metric space with the distance induced by the L^∞-norm

$$\|\sigma_1 - \sigma_2\|_\infty = \sup_{\omega \in \mathbb{R}} |\sigma_1(\omega) - \sigma_2(\omega)|.$$

We now consider the mapping

$$\phi : \Sigma_\omega(\mathbb{R}) \to \mathcal{L}(L^2(0, T)),$$
$$\sigma \to \phi_\sigma.$$

Then we have Theorem 1.6.

THEOREM 1.6. *The application ϕ is a contraction from $\Sigma_\omega(\mathbb{R})$ onto $\mathcal{L}(L^2(0, T))$:*

$$\forall (\sigma_1, \sigma_2) \in \Sigma_\omega(\mathbb{R}) \quad \|\phi_{\sigma_1} - \phi_{\sigma_2}\| \leqq \|\sigma_1 - \sigma_2\|_\infty.$$

Proof. The result is obvious for the difference $\|\hat{\phi}_{\sigma_1} - \hat{\phi}_{\sigma_2}\|$. The result then follows using causality and Plancherel's theorem. □

Now we can consider the case where σ is also a function of the space variable x. Then we consider the operator ψ_σ formally defined by

$$(\psi_\sigma u)(x, t) = \sigma_\infty(x) u(x, t) + \int_0^t \tilde{\sigma}_1(x, t - s) u(x, s) \, ds.$$

1.2. Study of the operator ψ_σ.

Notation for § § 1.2 *and* 2. $\Omega =]0, +\infty[$; $x \in \Omega$ is the space variable. $(\cdot, \cdot)$ and $\|\cdot\|$ denote the usual scalar product and norm in $L^2(\Omega)$.

1.2.1. The class $\Sigma(\Omega; \mathbb{R})$ of admissible polarization laws. We now introduce the following set of functions:

$$\Sigma(\Omega; \mathbb{R}) = \{\sigma : \Omega \times \mathbb{R} \to \mathbb{C} / (1.8) \text{ and } (1.9)\},$$

(1.8) $\exists (\sigma_\infty, \tilde{\sigma}_1) \in L^\infty(\Omega) \times L^\infty(\Omega; L^1(\mathbb{R}^+))$ such that $\sigma_\infty(x) \geqq (\sigma_\infty)_* > 0$ a.e. $x \in \Omega$, and $\forall \omega \in \mathbb{R} \quad \sigma(x, \omega) = \sigma_\infty(x) + 1/\sqrt{2\Pi} \int_0^{+\infty} \tilde{\sigma}_1(x, t)\, e^{-i\omega t} \, dt$ a.e. $x \in \Omega$,

(1.9) $$\exists \sigma_* \in \mathbb{R}_*^+ / \text{a.e. } (x, \omega) \in \Omega \times \mathbb{R} \quad \mathcal{R}e(\sigma(x, \omega)) \geqq \sigma_* > 0.$$

Properties of $\Sigma(\Omega; \mathbb{R})$. (1) $\Sigma(\Omega; \mathbb{R})$ is an open and convex subset of $L^\infty(\Omega; C_b^0(\mathbb{R}))$ for the topology of $L^\infty(\Omega \times \mathbb{R})$.

(2) For any σ in $\Sigma(\Omega;\mathbb{R})$, almost everywhere $x\in\Omega\ \sigma(x,\cdot)\in\Sigma_\omega(\mathbb{R})$, and almost everywhere $(x,\omega)\in\Omega\times\mathbb{R}\quad 0<\sigma_*\leqq\sigma(x,\omega)\leqq\sigma^*<+\infty$ (where $\sigma^*=\|\sigma_\infty\|_{L^\infty(\Omega)}+\|\tilde{\sigma}_1\|_{L^\infty(0,\,L^1(\mathbb{R}^+))}$).

1.2.2. Definition and properties of ψ_σ. We define the operator

$$\psi_\sigma: L^2(0,T;L^2(\Omega))\to L^2(0,T;L^2(\Omega)),$$
$$u(x,t)\to v(x,t)=(\psi_\sigma u)(x,t),$$
$$v(x,t)=\sigma_\infty(x)u(x,t)+\int_0^t\tilde{\sigma}_1(x,t-s)u(x,s)\,ds.$$

All the results concerning ψ_σ can be deduced from those obtained in § 1.1 for ϕ_σ. It suffices to note that

$$\text{a.e. } x\in\Omega\quad(\psi_\sigma u)(x,\cdot)=\phi_{\sigma(x,\cdot)}u(x,\cdot).$$

Then Theorems 1.4 and 1.5 lead to the two following results.

THEOREM 1.7. *ψ_σ is a linear, continuous, invertible operator in $L^2(0,T;L^2(\Omega))$. Its inverse ψ_σ^{-1} is continuous and we have* ($t\in[0,T]$)

$$\forall u L^2(0,T;L^2(\Omega))\quad \sigma_*\int_0^t\|u(s)\|^2\,ds\leqq\int_0^t(\psi_\sigma u(s),u(s))\,ds\leqq\sigma^*\int_0^t\|u(s)\|^2\,ds,$$

$$\|\psi_\sigma\|\leqq\sigma^*\quad\textit{and}\quad\|\psi_\sigma^{-1}\|\leqq(\sigma_*)^{-1}.$$

THEOREM 1.8. *For any u in $H^1(0,T;L^2(\Omega))$, we have*

$$\frac{d}{dt}(\psi_\sigma u)=\psi_\sigma\left(\frac{du}{dt}\right)+\tilde{\sigma}_1(t)\times u(x,0)\quad\textit{in }\mathcal{D}'(0,T;L^2(\Omega)).$$

Finally, let us give to $\Sigma(\Omega;\mathbb{R})$ a metric space structure with the following distance:

$$\|\sigma_1-\sigma_2\|_\infty=\sup_{(x,\omega)\in\Omega\times\mathbb{R}}|\sigma_1(x,\omega)-\sigma_2(x,\omega)|$$

and let us consider the mapping

$$\psi:\Sigma(\Omega;\mathbb{R})\to\mathcal{L}(L^2(0,T;L^2(\Omega)),$$
$$\sigma\to\psi_\sigma.$$

From Theorem 1.6, we deduce Theorem 1.9.

THEOREM 1.9. *The mapping ψ is a contraction from $\Sigma(\Omega,\mathbb{R})$ into $\mathcal{L}(L^2(0,T;L^2(\Omega)))$:*

$$\forall(\sigma_1,\sigma_2)\in\Sigma(\Omega;\mathbb{R})^2\quad\|\psi_{\sigma_1}-\psi_{\sigma_2}\|\leqq\|\sigma_1-\sigma_2\|_\infty.$$

1.3. Two particular subsets of $\Sigma(\Omega;\mathbb{R})$.

1.3.1. The subset $\Sigma_w(\Omega;\mathbb{R})$ of Warbourg's law. Warbourg's law is defined by

$$\sigma(x,\omega)=\sigma_0(x)\left(1+\lambda(x)\left(i\frac{\omega}{\omega_c(x)}\right)^{1/2}\Big/1+\left(i\frac{\omega}{\omega_c(x)}\right)^{1/2}\right)\qquad(\mathcal{R}\mathrm{e}\,(\mathbf{Z}^{1/2})>0)\tag{1.10}$$

where $\sigma_0(x)$, $\lambda(x)$, and $\omega_c(x)$ are strictly positive functions. Note that we can write

$$\sigma(x,\omega)=\sigma_\infty(x)+(1-\lambda(x))\left(\sigma_0(x)\Big/1+\left(i\frac{\omega}{\omega_c(x)}\right)^{1/2}\right),$$

$$\sigma_\infty(x)=\lambda(x)\sigma_0(x)=\lim_{\omega\to\infty}\sigma(x,\omega).$$

THEOREM 1.10. *Under the following assumptions*:

$$0<(\sigma_0)_*\leqq\sigma_0(x)\leqq(\sigma_0)^* \quad a.e.\ x\in\Omega,$$

$$1\leqq\lambda_*\leqq\lambda(x)\leqq\lambda^* \quad a.e.\ x\in\Omega,$$

$$0<(\omega_c)_*\leqq\omega_c(x)\leqq(\omega_c)^* \quad a.e.\ x\in\Omega,$$

the law σ *defined by* (1.10) *belongs to* $\Sigma(\Omega;\mathbb{R})$.

Proof. We have to check that σ satisfies both properties (1.8) and (1.9). To prove this result we will use Lemma 1.1.

LEMMA 1.1. *Let* $f(t)$ *be the real and causal function defined by*

$$f(t)=\frac{1}{\sqrt{\Pi}}\left(\frac{1}{\sqrt{t}}-2\,e^{t}\int_{\sqrt{t}}^{+\infty}e^{-s^2}\,ds\right) \quad \text{if } t>0,$$

$$f(t)=0 \quad \text{if } t\leqq 0.$$

Then, f *is positive, belongs to* $L^1(\mathbb{R}^+)$, *and satisfies*

$$\|f\|_{L^1(\mathbb{R}^+)}=1.$$

Moreover its Fourier transform is given by

$$F(\omega)=(\mathscr{F}f)(\omega)=\frac{1}{1+(i\omega)^{1/2}}.$$

Proof of Lemma 1.1. (i) The calculation of $f(t)$ (as inverse Fourier transform of $\omega\to(1+(i\omega)^{1/2})^{-1}$) can be found in [4].

(ii) The positivity of $f(t)$ is a consequence of the following inequality:

$$\int_{\sqrt{t}}^{+\infty}e^{-s^2}\,ds=\int_{\sqrt{t}}^{+\infty}\frac{s\,e^{-s^2}}{s}\,ds\leqq\frac{1}{\sqrt{t}}\int_{\sqrt{t}}^{+\infty}s\,e^{-s^2}\,ds=\frac{1}{2\sqrt{t}}\,e^{-t}.$$

(iii) As $\lim_{t\to 0}e^{t}\int_{\sqrt{t}}^{+\infty}e^{-s^2}\,ds<+\infty$, we have

$$f(t)=\frac{1}{\sqrt{\Pi t}}+O(1) \quad \text{when } t\to 0.$$

(iv) A double integration by parts (see [4] for details) leads to the following identity:

$$\int_{\sqrt{t}}^{+\infty}e^{-s^2}\,ds=\frac{1}{2}\frac{e^{-t}}{\sqrt{t}}-\frac{1}{4}\frac{e^{-t}}{t^{3/2}}+\frac{3}{4}\int_{\sqrt{t}}^{+\infty}\frac{e^{-s^2}}{s^4}\,ds$$

from which we easily deduce that

$$f(t)=\frac{1}{2\sqrt{\Pi}}\frac{e^{-t}}{t^{3/2}}\left(1+O\left(\frac{1}{t}\right)\right) \quad \text{when } t\to+\infty.$$

Stems (iii) and (iv) prove that f belongs to $L^1(\mathbb{R}^+)$ and as f is positive

$$\|f\|_{L^1(\mathbb{R}^+)}=f(0)=1. \qquad \square$$

By Lemma 1.1, we can write

$$\sigma(x,\omega)=\sigma_\infty(x)+\frac{1}{\sqrt{2\Pi}}\int_0^\infty\tilde{\sigma}_1(x,t)\,e^{-i\omega t}\,dt$$

with

$$\tilde{\sigma}_1(x, t) = (1-\lambda(x))\sigma_0(x) f(\omega_c(x)t).$$

This proves that (1.8) is satisfied. Now, a simple calculation gives

$$\mathcal{R}\mathrm{e}(\sigma(x, \omega)) = \lambda(x)\sigma_0(x) + (1-\lambda(x))\sigma_0(x)\psi\left(\frac{\omega}{\omega_c(x)}\right)$$

where

$$\psi(y) = \frac{1+y\sqrt{2}/2}{1+y\sqrt{2}+y^2}.$$

It is easy to check that

$$\forall y \in \mathbb{R} \quad 0 < \psi(y) \leqq 1.$$

Then, as $1-\lambda(x) < 0$, we obtain

$$\mathcal{R}\mathrm{e}(\sigma(x, \omega)) \geqq \lambda(x)\sigma_0(x) + (1-\lambda(x))\sigma_0(x) \geqq (\sigma_0)_* > 0,$$

which proves (1.9) and the theorem. □

We will denote by $\Sigma_w(\Omega; \mathbb{R})$ the subset of Warbourg's law defined by (1.10) with the assumptions of Theorem 1.10.

1.3.2. The subset $\Sigma_a(\Omega; \mathbb{R})$ of approximate laws. We define an approximate polarization law by the formula

$$\sigma(x, \omega) = \sigma_\infty(x) + \sum_{k=1}^{n} \frac{b_k(x)}{1 + i a_k(x)\omega} \tag{1.11}$$

with the following assumptions (for each $k = 1, 2, \cdots, n$):

$$0 < (a_k)_* \leqq a_k(x) \leqq (a_k)^* < +\infty, \tag{1.12}_1$$

$$b_k(x) \leqq (b_k)^* < +\infty, \tag{1.12}_2$$

$$\inf_{(x, \omega)} \left\{ \sigma_\infty(x) + \sum_{k=1}^{n} \frac{b_k(x)}{1 + a_k(x)^2\omega^2} \right\} > 0. \tag{1.12}_3$$

THEOREM 1.11. *Under assumptions* (1.12), *the function* σ *defined by* (1.11) *belongs to* $\Sigma(\Omega; \mathbb{R})$.

Proof. We have the identity

$$\sigma(x, \omega) = \sigma_\infty(x) + \frac{1}{\sqrt{2\Pi}} \int_0^{+\infty} \tilde{\sigma}_1(x, t)\, e^{-i\omega t}\, dt$$

where

$$\tilde{\sigma}_1(x, t) = \sum_{k=1}^{n} \frac{b_k(x)}{a_k(x)} e^{-t/a_k(x)},$$

which proves (1.8). Moreover, (1.9) is nothing but assumption $(1.12)_3$ (which is true if $b_k(x) \geqq 0$ for each k). □

We will denote by $\Sigma_a(\Omega; \mathbb{R})$ the set of approximate laws defined by (1.11). Of course the main interest of $\Sigma_a(\Omega; \mathbb{R})$ lies in the fact that, for σ in $\Sigma_a(\Omega; \mathbb{R})$, the corresponding operator ψ_σ can be defined with the help of ordinary differential operators. More precisely we have the following result.

THEOREM 1.12. *The operator ψ_σ associated with the law* (1.11) *is defined by*

$$\forall u \in L^2(0, T; L^2(\Omega)) \quad (\psi_\sigma u)(x, t) = \sigma_\infty(x, t) + \sum_{k=1}^{n} f_k(x, t)$$

where each function $f_k(x, t)$ is the unique solution of

$$a_k(x)\frac{\partial f_k}{\partial x}(x, t) + f_k(x, t) = b_k(x)u(x, t), \qquad f_k(x, 0) = 0.$$

2. Mathematical analysis of system (1.1). Let σ be an element of $\Sigma(\Omega; \mathbb{R})$. We consider the following initial boundary value problem:

$$\text{(2.1)} \quad \begin{cases} \mu\dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, & (x, t) \in \Omega \times \mathbb{R}^+, \\ \psi_\sigma v + \dfrac{\partial u}{\partial x} = 0, & (x, t) \in \Omega \times \mathbb{R}^+, \\ u(x, 0) = u_0(x), & x \in \Omega, \\ v(0, t) = \phi(t), & t \in \mathbb{R}^+. \end{cases}$$

In view of the study of continuity with respect to σ, we are going to deal with the nonhomogeneous problem:

$$\text{(2.2)} \quad \begin{cases} \mu\dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, & (x, t) \in \Omega \times \mathbb{R}^+, \\ \psi_\sigma v + \dfrac{\partial u}{\partial x} = g, & (x, t) \in \Omega \times \mathbb{R}^+, \\ u(x, 0) = u_0(x), & x \in \Omega, \\ v(0, t) = \phi(t), & t \in \mathbb{R}^+, \end{cases}$$

with the following assumptions

$$\text{(2.3)} \qquad u_0 \in L^2(\Omega), \quad \phi \in L^2(0, T), \quad g \in L^2(0, T; L^2(\Omega)).$$

2.1. Existence, uniqueness, and regularity results.

2.1.1. Notation and functional spaces. We introduce the Banach space

$$\mathcal{H}(0, T) = W(0, T) \times L^2(0, T; L^2(\Omega))$$

where $W(0, T) = \{u \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega)) / du/dt \in L^2(0, T; H^{-1}(\Omega))\}$ with the following norms:

$$\|(u, v)\|_{\mathcal{H}(0,T)} = \|u\|_{W(0,T)} + \|v\|_{L^2(0,T;L^2(\Omega))},$$

$$\|u\|_{W(0,T)} = \|u\|_{L^\infty(0,T;L^2(\Omega))} + \|u\|_{L^2(0,T;H^1(\Omega))} + \left\|\frac{du}{dt}\right\|_{L^2(0,T;H^{-1}(\Omega))}.$$

We are going to look for solutions of (2.2) in the space $\mathcal{H}$. Note that for (u, v) in $\mathcal{H}$, it is not possible to give a meaning (in the classical sense) to the boundary condition $v(0, t) = \phi(t)$

Nevertheless, it is possible to give a "weak" meaning to this condition via the following definition.

DEFINITION. An element (u, v) in $\mathcal{H}(0, T)$ is a weak solution of (2.2) if and only if

$$\forall u^* \in H^1(\Omega) \quad \mu \frac{d}{dt}(u(t), u^*) - \left(v(t), \frac{du^*}{dx}\right) + u^*(0)\phi(t) = 0 \quad \text{in } \mathcal{D}'(0, T),$$

$$\forall v^* \in H^1(\Omega) \quad (\psi_\sigma v(t), v^*) + \left(\frac{\partial u}{\partial x}(t), v^*\right) = (g(t), v^*) \quad \text{in } \mathcal{D}'(0, T),$$

$$u(0) = u_0.$$

Remarks. (1) If (u, v) is a weak solution we have

$$\mu \frac{\partial u}{\partial t} - \frac{\partial v}{\partial x} = 0 \quad \text{in } L^2(0, T; H^{-1}(\Omega)),$$

$$\psi_\sigma v - \frac{\partial u}{\partial x} = g \quad \text{in } L^2(0, T; L^2(\Omega)).$$

(2) As soon as v belongs to $L^2(0, T; H^1(\Omega))$ we have

$$v(0, t) = \phi(t) \quad \text{in } L^2(0, T).$$

(3) The initial condition $u(0) = u_0$ makes sense since $W(0, T) \hookrightarrow C^0(0, T; H^{-1}(\Omega))$.

2.1.2. The existence and uniqueness theorem.

THEOREM 2.1. *Under assumption* (2.3), *problem* (2.2) *admits a unique weak solution* (u, v) *in* $\mathcal{H}(0, T)$. *Moreover, we have the following a priori estimates*:

$$\|u\|_{L^\infty(0,T;L^2(\Omega))} \leqq \|u_0\| + C\left(\|g\| + \|\phi\| + \int_0^T |\phi(t)|\, dt\right),$$

$$\|v\|_{L^2(0,T;L^2(\Omega))} \leqq C\left(\|u_0\| + \|g\| + \|\phi\| + \int_0^T |\phi(t)|\, dt\right),$$

$$\left\|\frac{\partial u}{\partial x}\right\|_{L^2(0,T;L^2(\Omega))} \leqq C\left(\|u_0\| + \|g\| + \|\phi\| + \int_0^T |\phi(t)|\, dt\right),$$

$$\left\|\frac{\partial u}{\partial t}\right\|_{L^2(0,T;H^{-1}(\Omega))} \leqq C\left(\|u_0\| + \|g\| + \|\phi\| + \int_0^T |\phi(t)|\, dt\right),$$

where the constants C *only depend on* μ, σ_*, *and* σ^* *and where we have set*

$$\|\phi\| = \|\phi\|_{L^2(0,T)}, \qquad \|g\| = \|g\|_{L^2(0,T;L^2(\Omega))}.$$

Proof of the theorem. (1) *The a priori estimates.*

$$\mu \frac{\partial u}{\partial t} + \frac{\partial v}{\partial x} = 0, \tag{2.4}_1$$

$$\psi_\sigma v + \frac{\partial u}{\partial x} = g. \tag{2.4}_2$$

Multiply $(2.4)_1$ by u, $(2.4)_2$ by v, add the two equations, and integrate in space. Integrating by parts, we easily obtain

$$\frac{\mu}{2}\frac{d}{dt}\|u(s)\|^2 + (\psi_\sigma v(s), v(s)) = \phi(s)u(0, s) + (g(s), v(s)).$$

Now, integrating in time between 0 and t ($t \in [0, T]$) gives the following identity:

$$\frac{\mu}{2}\|u(t)\|^2 + \int_0^t (\psi_\sigma v(s), v(s))\, ds = \frac{\mu}{2}\|u_0\|^2 + \int_0^t (g(s), v(s))\, ds + \int_0^t \phi(s) u(0, s)\, ds.$$

We use the ellipticity of the operator ψ_σ (Theorem 1.7) to obtain

$$\frac{\mu}{2}\|u(t)\|^2 + \sigma_* \int_0^t \|v(s)\|^2\, ds \leqq \frac{\mu}{2}\|u_0\|^2 + \int_0^t \|g(s)\|\, \|v(s)\|\, ds + \int_0^t |\phi(s)|\, |u(0, s)|\, ds.$$

The trace theorem furnishes the following estimate:

$$\begin{aligned} |u(0, s)| &\leqq \|u(s)\| + \left\|\frac{\partial u}{\partial x}(s)\right\| \\ &\leqq \|u(s)\| + \|g(s)\| + \|\psi_\sigma u(s)\| \qquad ((2.4)_2). \end{aligned}$$

Then, standard techniques (Young inequalities and a generalized Gronwall inequality) together with the continuity property of ψ_σ lead to the estimates on $\|u\|_{L^\infty(0,T;L^2(\Omega))}$ and $\|v\|_{L^2(0,T;L^2(\Omega)}$.

The estimate on $\partial u/\partial x$ stems from $(2.4)_1$ and from the fact that ψ_σ belongs to $\mathscr{L}(L^2(0, T; L^2(\Omega)))$, the estimate on $\partial u/\partial t$ from $(2.4)_4$, and from the fact that $\partial/\partial x \in \mathscr{L}(L^2(\Omega); H^{-1}(\Omega))$.

Clearly the uniqueness result is a consequence of the a priori estimates.

(2) *Existence proof by Galerkin's method.* Let $w_1, \cdots, w_n, \cdots$ be a Hilbert space basis of $H^1(\Omega)$. We construct an approximate solution (u_n, v_n) in the following way:

(2.5) Find $(u_n(t), v_n(t)) : [0, T] \to V_n \times V_n$ (where $V_n = \operatorname{span}[w_1, w_2, \cdots, w_n]$) such that ($g_n \in L^2(0, T; V_n)$)

$$\forall u^* \in V_n \qquad \mu \frac{d}{dt}(u_n(t), u^*) - \left(u_n(t), \frac{\partial u^*}{\partial x}\right) + u^*(0)\Phi(t) = 0,$$

$$\forall u^* \in V_n \qquad (\Psi_\sigma v_n(t), v^*) + \left(\frac{\partial u_n}{\partial x}(t), v^*\right) = (g_n(t), v^*),$$

$$u_n(0) = u_{0,n} \in V_n.$$

Decomposing the approximate solution on the basis $\{w_1, w_2, \cdots, w_n\}$ it is easy to see that solving P_n is equivalent to solving the following integrodifferential system in $\mathbb{R}^n$:

$$\begin{aligned} &A_n \frac{dU_n}{dt}(t) - B_n V_n(t) + \phi_n(t) = 0 \quad \text{in } \mathbb{R}^n, \\ &\Sigma_\infty^n V_n(t) + \int_0^t \Sigma_1^n(t - s) V_n(s)\, ds + {}^t B_n U_n(t) = G_n(t) \quad \text{in } \mathbb{R}^n, \\ &U_n(0) = U_{0,n} \in \mathbb{R}^n, \end{aligned} \tag{2.6}$$

where the different matrices occurring in (2.6) are the following:

$$(A_n)_{ij} = \mu(w_i, w_j) \quad \text{(symmetric positive definite)},$$

$$(B_n)_{ij} = \left(\frac{dw_i}{dx}, w_j\right) \quad \text{(antisymmetric)},$$

$$(\Sigma_\infty^n)_{ij} = \int_\Omega \sigma_\infty(x) w_i(x) w_j(x)\, dx \quad \text{(symmetric positive definite)},$$

$$(\Sigma_1^n)_{ij}(t) = \int_\Omega \tilde{\sigma}_1(x, t) w_i(x) w_j(x)\, dx \quad \text{(belongs to } L^1(0, T; \mathscr{L}(\mathbb{R}^n))).$$

The only nonstandard point is the following lemma.

LEMMA 2.1. *The approximate problem* (2.5) *admits one unique solution*

$$(u_n, v_n) \in H^1(0, T; V_n) \times C^0(0, T; V_n).$$

Proof. We write system (2.6) in its integral form:

$$A_n(U_n(t) - U_0^n) - B_n \int_0^t V_n(s)\, ds + \int_0^t \phi(s)\, ds = 0,$$

$$\Sigma_\infty^n V_n(t) + {}^tB_n U_n(t) + \int_0^t \Sigma_1^n(t-s) V_n(s)\, ds = \int_0^t G_n(s)\, ds,$$

which is strictly equivalent to

$$\Sigma_\infty^n V_n(t) = \int_0^t G_n(s) ds + \int_0^t \Sigma_1^n(t-s) V_n(s)\, ds - {}^tB_n\left(U_0 + A_n^{-1} B_n \int_0^t V_n(s)\, ds - A_n^{-1} \int_0^t \phi_n(s)\, ds \right), \tag{2.7}$$

$$U_n(t) = U_0^n + A_n^{-1} B_n \int_0^t V_n(s)\, ds - A_n^{-1} \int_0^t \phi_n(s)\, ds. \tag{2.8}$$

Let us introduce, in the Banach space $C^0(0, T_n: \mathbb{R})$ $(0 < T_n \leqq T)$

$$\Phi_n : C^0(0, T_n; \mathbb{R}^n) \to C^0(0, T_n; \mathbb{R}^n),$$

$$V(t) \to \Phi_n V(t),$$

$$\Phi_n V(t) = (\Sigma_\infty^n)^{-1} \left\{ \int_0^t G_n(s)\, ds - \int_0^t \Sigma_1^n(t-s) V(s)\, ds \right\} - (\Sigma_\infty^n)^{-1}\, {}^tB_n \left\{ U_0^n + A_n^{-1} B_n \int_0^t V(s)\, ds - A_n^{-1} \int_0^t \phi_n(s)\, ds \right\}.$$

A simple calculation shows that, for any (V_1, V_2) in $C^0(0, T_n; V_n)$,

$$\|\Phi_n V_2 - \Phi_n V_1\| \leqq |(\Sigma_\infty^n)^{-1}| \left(\int_0^{T_n} |(\Sigma_1^n(t)|\, dt) + |(\Sigma_\infty^n)^{-1}\, {}^tB_n A_n^{-1} B_n|\, T_n \right) \| V_2 - V_1 \|_{L^\infty(0, T_n)}$$

(where $|\cdot|$ denotes a convenient norm in $\mathscr{L}(\mathbb{R}^n)$). As $\Sigma_1(t)$ belongs to $L^1(0, T; \mathscr{L}(\mathbb{R}^n))$, we can choose T_n small enough so that

$$\operatorname*{Max}_{a \in \mathbb{R}} \left\{ |(\Sigma_\infty^n)^{-1}| \int_a^{a+T_n} |\Sigma_1^n(s)|\, ds + T_n |(\Sigma_\infty^n)^{-1}\, {}^tB_n A_n^{-1} B_n| \right\} < \frac{1}{2}.$$

Then for all $(V_1, V_2) \in C^0(0, T_n; \mathbb{R})^2$, $\|\Phi_n V_1 - \Phi_n V_2\|_{L^\infty(0, T_n)} \leqq \frac{1}{2} \| V_2 - V_1 \|_{L^\infty(0, T_n)}$.

So, using the Contraction Mapping Theorem, we show that there exists a unique V_n in $C^0(0, T_n; \mathbb{R}^n)$ such that $\Phi_n V_n = V_n (\Leftrightarrow (2.7))$. Of course, we can iterate the process on the interval $[T_n, 2T_n]$, $[2T_n, 3T_n]$, $\cdots$ and construct a solution in $C^0(0, T; \mathbb{R}^n)$. U_n is then obtained by (2.8), which shows that U_n belongs to $H^1(0, T; V_n)$. □

It is then standard to prove that, if we choose $u_{0,n}$ and g_n such that

$$(u_{0,n}, g_n) \to (u_0, g) \text{ in } L^2(\Omega) \times L^2(0, T; L^2(\Omega)),$$

the sequence (u_n, v_n) converges (weakly in $L^2(0, T; L^2(\Omega))^2$) to a solution (u, v) of (2.2) (see, for example, [8]). □

2.1.3. Regularity results. Let us consider two different groups of assumptions:

$$\begin{cases} \phi \in H^1(0, T), \\ u_0 = 0, \\ \phi(0) = 0 \qquad \text{(compatibility condition)}, \end{cases} \tag{2.9}$$

$$\begin{cases} (\phi, \tilde{\sigma}_1) \in H^1(0, T) \times L^2(0, T), \\ \dfrac{\partial}{\partial x}\left(\sigma_\infty^{-1} \dfrac{\partial u_0}{\partial x}\right) \in L^2(\Omega), \\ \phi(0) = \left(\sigma_\infty^{-1} \dfrac{\partial u_0}{\partial x}\right)(0) \qquad \text{(compatibility condition)}. \end{cases} \tag{2.10}$$

Remark. For $\sigma \in \Sigma_a(\Omega; \mathbb{R})$, $\tilde{\sigma}_1$ belongs to $L^2(0, T)$; however, for σ in $\Sigma_w(\Omega; \mathbb{R})$, $\tilde{\sigma}_1$ does not belong to $L^2(0, T)$.

We can show (see [4] for details) the following theorem.

THEOREM 2.2. *Under assumptions* (2.9) *or* (2.10), *the unique solution* (u, v) *of* (2.1) *satisfies*

$$u \in W^{1,\infty}(0, T; L^2(\Omega)) \cap H^1(0, T; H^1(\Omega)) \cap H^2(0, T; H^{-1}(\Omega)),$$

$$v \in L^2(0, T; H^1(\Omega)) \cap H^1(0, T; L^2(\Omega)),$$

and therefore, we have the following equality:

$$v(0, t) = \phi(t) \quad \text{in } L^2(0, T).$$

Comments. (1) This result can be easily obtained by considering the system satisfied by $(\partial u/\partial t, \partial v/\partial t)$. (We use Theorem 1.8 to obtain it.)

(2) The assumptions (2.9) or (2.10) are necessary to obtain the time regularity up to the boundary, but they are not necessary to obtain "interior" time regularity. There is, as for the diffusion equation, a regularizing effect as soon as $x > 0$, which is due to the property

$$|\sigma(x, \omega)| \geqq \sigma_* > 0.$$

(See [4] for a more precise result.)

(3) It is not clear whether there exists a maximum principle for this system.

2.2. Continuity of the solution of (2.1) with respect to the polarization law σ. Using Theorem 2.1 (with $g = 0$), we can construct the mapping

$$\mathcal{U}_T: L^2(\Omega) \times L^2(0, T) \times \Sigma(\Omega; \mathbb{R}) \to \mathcal{H}(0, T),$$

$$(u_0, \phi, \sigma) \to (u, v),$$

(u, v) is the unique solution of (2.1).

The mapping $(u_0, \phi) \to \mathcal{U}_T(\sigma, u_0, \phi)$ is linear and continuous (Theorem 2.1).

The mapping $\sigma \to \mathcal{U}_T(\sigma, u_0, \phi)$ is nonlinear.

THEOREM 2.3. *For given* (ϕ, u_0), *the mapping* $\sigma \to \mathcal{U}_T(\sigma, u_0, \phi)$ *is locally Lipschitz continuous from* $\Sigma(\Omega, \mathbb{R})$ *into* $\mathcal{H}(0, T)$:

$$\forall (\sigma_1, \sigma_2) \in \Sigma(\Omega; \mathbb{R}), \quad \exists C(\mu, \sigma_1, \sigma_2, u_0, \phi, T) \text{ such that}$$

$$\|\mathcal{U}_T(\sigma_2, u_0, \phi) - \mathcal{U}_T(\sigma_1, u_0, \phi)\|_{\mathcal{H}(0,T)} \leqq C \|\sigma_1 - \sigma_2\|_\infty.$$

Proof. Let $(u_1, v_1) = \mathcal{U}_T(\sigma_1, u_0, \phi)$, $(u_2, v_2) = \mathcal{U}_T(\sigma_2, u_0, \phi)$, and $(\tilde{u}, \tilde{v}) = (u_1 - u_2, v_1 - v_2)$. Then, $(\tilde{u}, \tilde{v})$ is the solution of the following system:

$$\mu \frac{\partial \tilde{u}}{\partial t} + \frac{\partial \tilde{v}}{\partial x} = 0, \qquad \psi_{\sigma_1} \tilde{v} + \frac{\partial \tilde{u}}{\partial x} = (\psi_{\sigma_2} - \psi_{\sigma_1}) v_2,$$

$$\tilde{u}(x, 0) = 0, \qquad \tilde{v}(0, t) = 0.$$

Using Theorem 2.1 with $(u_0, \phi, g) = (0, 0, (\psi_{\sigma_2} - \psi_{\sigma_1}) v_2)$, we can obtain the estimate

$$\|(\tilde{u}, \tilde{v})\|_{\mathcal{H}(0,T)} \leqq C(\mu, \sigma_1) \|(\psi_{\sigma_2} - \psi_{\sigma_1}) v_2\|_{L^2(0,T;L^2(\Omega))}.$$

However, by Theorem 1.9 we know that

$$\|(\psi_{\sigma_2} - \psi_{\sigma_1}) v_2\|_{L^2(0,T;L^2(\Omega))} \leqq \|\sigma_2 - \sigma_1\|_\infty \|v_2\|_{L^2(0,\, T;\, L^2(\Omega))}$$

and the result follows with

$$C(\mu, \sigma_1, \sigma_2, u_0, \phi, T) = C_1(\mu, \sigma_1) \|v_2\|_{L^2(0,\, T;\, L^2(\Omega))}. \qquad \square$$

Remark. In view of the solution of the identification problem, it would also be interesting to study the differentiability of the mapping $\sigma \to \mathcal{U}_T(\sigma, u_0, \phi)$.

3. Convergence result and error estimate.

3.1. A general approximation result. Let us recall that the numerical method we propose for solving problem (1.1) begins by substituting for the exact polarization law σ an approximate law σ_a belonging to $\Sigma_a(\Omega; \mathbb{R})$ (cf. § 1.3.2), which reduces the approximate system to a system of two linear partial differential equations coupled to n ordinary differential equations (cf. Theorem 1.12 and (0.4)).

From a theoretical point of view, the continuity result obtained in § 2.2 justifies this approach for any function σ that belongs to the closure of $\Sigma_a(\Omega; \mathbb{R})$ in $\Sigma(\Omega; \mathbb{R})$. Indeed, by assuming that

$$\sigma \in \overline{\Sigma_a(\Omega; \mathbb{R})}^{\Sigma(\Omega;\mathbb{R})},$$

there exists a sequence σ_n in $\Sigma_a(\Omega; \mathbb{R})$ such that

$$\sigma_n \to \sigma \quad \text{in } \Sigma(\Omega; \mathbb{R}).$$

Then, by Theorem 2.3, the approximate solution (u_n, v_n) of problem (2.1) associated with σ_n converges in $\mathcal{H}(0, T)$ to the exact solution (u, v) associated with σ and we have the error estimate (Theorem 2.3):

$$\|(u_n, v_n) - (u, v)\|_{\mathcal{H}(0,T)} \leqq C(\sigma) \|\sigma_n - \sigma\|_\infty.$$

The aim of this section is to prove that such a sequence exists for any Warbourg law, in other words,

$$\Sigma_w(\Omega; \mathbb{R}) \subset \overline{\Sigma_a(\Omega; \mathbb{R})}^{\Sigma(\Omega;\mathbb{R})}.$$

3.2. Approximation of Warbourg's law by rational fractions. First we notice that if we make the approximation

$$f(\omega) = \frac{1}{1 + (i\omega)^{1/2}} \simeq f_n(i\omega) = \sum_{k=1}^{n} \frac{\beta_n^k}{1 + i\alpha_n^k \omega},$$

then Warbourg's law (cf. § 1.3.1),

$$\sigma(x, \omega) = \sigma_\infty(x) + [\lambda(x) - 1] \sigma_0(x) f\left(i \frac{\omega}{\omega_c(x)}\right),$$

can be approximated by

$$\sigma_a(x,\omega)=\sigma_\infty(x)+[\lambda(x)-1]\sigma_0(x)f_n\left(i\frac{\omega}{\omega_c(x)}\right),$$

which can be written in the form (1.11) with

$$a_k(x)=\alpha_n^k\omega_c(x)^{-1},\qquad b_k(x)=\beta_n^k\sigma_0(x)(\lambda(x)-1).$$

To approximate $f(z)=(1+z^{1/2})^{-1}$ by rational fractions, a rather classical method (see [1] or [7] for other applications) consists in using continued fractions (see [9] for the general theory)

Remarking that $f(z)$ is a fixed point of an homographic function,

$$f(z)=\frac{1}{2+(z-1)f(z)}\quad \forall z\notin\mathbb{R}^-, \tag{3.1}$$

we apply the process of successive approximations to define $f_n^*(z)$ by

$$f_{n+1}^*(z)=\frac{1}{2+(z-1)f_n^*(z)},\qquad f_0^*(z)=0. \tag{3.2}$$

LEMMA 3.1. *For each n in* $\mathbb{N}$, $f_n^*(z)$ *is a rational fraction.*

Proof. By induction it is easy to check that

$$f_{2n}^*(z)=\frac{P_{n-1}(x)}{Q_n(z)},\qquad f_{2n+1}^*(z)=\frac{Q_n(z)}{P_n(z)},$$

where (P_n, Q_n) are the sequences of polynomials defined by the following relations:

$$P_{n+1}(z)=(z-1)P_n(z)+2Q_n(z),$$
$$Q_{n+1}(z)=(z-1)Q_n(z)+2P_n(z),$$
$$P_0(z)=0,\qquad Q_0(z)=2,$$

which show that $d^0P_n=d^0Q_n=n$. □

LEMMA 3.2. *The sequence* $f_n^*(z)$ *is given by*

$$\forall z\in\mathbb{C}\backslash\mathbb{R}^-\quad f_n^*(z)=f(z)\left\{\left(1-\left(\frac{1-z^{1/2}}{1+z^{1/2}}\right)^n\right)\Big/\left(1-\left(\frac{1-z^{1/2}}{1+z^{1/2}}\right)^{n+1}\right)\right\},$$

$$f_n^*(0)=1-\frac{1}{n+1}.$$

Proof. Let us set $f(z)=(1+\sqrt{z})^{-1}$ and $g(z)=(1+\sqrt{z})^{-1}$, which are the two solutions of $y=1/(2+(z-1)y)$. We easily check that

$$\frac{f_{n+1}^*(z)-f(z)}{f_{n+1}^*(z)-g(z)}=\frac{1-\sqrt{z}}{1-\sqrt{z}}\frac{f_n^*(z)-f(z)}{f_n^*(z)-g(z)},$$

so that

$$\frac{f_n^*(z)-f(z)}{f_n^*(z)-g(z)}=\left(\frac{1-\sqrt{z}}{1-\sqrt{z}}\right)^n\frac{f(z)}{g(z)},$$

which leads to Lemma 3.3. □

LEMMA 3.3. *For each z in* $\mathbb{C}\backslash\mathbb{R}^-$,

$$\lim_{n\to+\infty} f_n^*(z)=f(z).$$

Moreover, for each compact subset K of $\mathbb{C}\backslash\mathbb{R}^-$, *there exists a constant* $\alpha(K)\in[0,1]$ *such that*

$$|f_n^*(z)-f(z)|\leqq C\alpha(K)^n.$$

Proof. From Lemma 3.2, we deduce the equality

$$f(z)-f_n^*(z)=\frac{2\sqrt{z}}{1+\sqrt{z}}\left(\left(\frac{1-\sqrt{z}}{1+\sqrt{z}}\right)^n\Big/\left(1-\left(\frac{1-\sqrt{z}}{1+\sqrt{z}}\right)^{n+1}\right)\right);$$

then it suffices to remark that, as $\mathcal{R}e(\sqrt{z})>0$, the function $\phi(z)=(1-\sqrt{z})/(1+\sqrt{z})$ maps $\mathbb{C}\backslash i\mathbb{R}^-$ to the unit open disk $D=\{z\in\mathbb{C}/|z|<1\}$. □

We can now state the first important result of this section.

THEOREM 3.1. *For any n, the rational fractions* $f_n^*(z)$ *admit the following expansions*:

(i) $$f_{2n}^*(z)=\frac{2}{2n+1}\sum_{k=1}^{n}1\Big/\left(1+z\,\mathrm{cotg}^2\frac{k\Pi}{2n+1}\right).$$

(ii) $$f_{2n+1}^*(z)=\frac{2}{2n+1}+\frac{2}{2n+2}\sum_{k=1}^{n}1\Big/\left(1+z\,\mathrm{cotg}^2\frac{k\Pi}{2n+2}\right).$$

Proof. Let us give the proof for $f_{2n}^*(z)$.

(1) *Determination of the poles of* f_{2n}^*. Using Lemma 3.2, the poles of $f_{2n}^*(z)$ are the solution of

$$\left(\frac{1-\sqrt{z}}{1+\sqrt{z}}\right)^{2n+1}=1 \quad\text{and}\quad \left(\frac{1-\sqrt{z}}{1+\sqrt{z}}\right)\neq 1.$$

$$\Leftrightarrow((1-\sqrt{z})/(1+\sqrt{z}))=e^{2ik\Pi/(2n+1)},\qquad k=1,2,\cdots,2n.$$

$$\Leftrightarrow z=-\mathrm{tg}^2(k\Pi/(2n+1)),\qquad k=1,2,\cdots,2n.$$

(2) Using Lemma 3.1, we know that $f_{2n}^*(z)$ has an expansion in the form

$$f_{2n}^*(z)=\sum_{k=1}^{n}\frac{r_{2n}^k}{z+\mathrm{tg}^2(\mathrm{k}\Pi/(2\mathrm{n}+1))}.$$

As each pole $z_{2n}^k=-\mathrm{tg}^2(k\Pi/(2n+1))$ is simple, we have $r_{2n}^k=\{(1/f_{2n}^*)'(z_{2n}^k)\}^{-1}$. Then, if we use the expression of $f_n^*(z)$ given in Lemma 3.2, it is rather easy to obtain

$$r_{2n}^k=\frac{2}{2n+1}\mathrm{tg}^2\left(\frac{k\Pi}{2n+1}\right)$$

which completes the proof. □

We now consider the sequence $f_n(z)$ defined by

$$\forall n\in\mathbb{N}\quad f_n(z)=f_{2n+1}^*(z),$$

for which we can establish the following approximation result.

THEOREM 3.2. *For any integer n in* $\mathbb{N}$, *we have the equality*

$$\sup_{\mathcal{R}e(z)\geqq\phi}|f_n(z)-f(z)|=\frac{1}{2(n+1)}.$$

Proof of the theorem. (1) A first calculation proves that for any $z\neq 0$, we have the identity

$$f_n\left(\frac{1}{z}\right)-f\left(\frac{1}{z}\right)=f(z)-f_n(z).$$

(2) Now we can restrict the study to $|z|\leqq 1$.

From Lemma 3.2, we deduce the identity

$$f_n(z)-f(z)=\frac{1}{1+\sqrt{z}}\frac{1-\psi(z)}{1-\psi(z)^{2n+2}} \quad \text{where } \psi(z)=\frac{1+\sqrt{z}}{1-\sqrt{z}}.$$

Thus we can also write

$$f_n(z)-f(z)=\frac{1}{(1+\sqrt{z})(1+\psi(z))}\left\{\prod_{\substack{k=1\\k\neq n+1}}^{2n+1}(\psi(z)-e^{ik\pi/(n+1)})\right\}^{-1}.$$

A simple calculation gives

$$\psi(z)-e^{ik\pi/(n+1)}=\left(\sqrt{z}-i\,\mathrm{tg}\frac{k\Pi}{2(n+1)}\right),$$

from which we deduce

$$f_n(z)-f(z)=\frac{(1-\sqrt{z})^{2n}}{(1+\sqrt{z})(1+\psi(z))}\,\delta_n(z), \tag{3.4}$$

$$\delta_n(z)=\left\{\prod_{\substack{k=1\\k\neq n+1}}^{2n+1}(1+e^{ik\Pi/(n+1)})\left(\sqrt{z}-i\,\mathrm{tg}\frac{k\Pi}{2(n+1)}\right)\right\}^{-1}. \tag{3.5}$$

A rearrangement of the terms of the product leads to

$$\delta_n(z)=\left(\prod_{k=1}^{n}|1+e^{ik\Pi/(n+1)}|^2\left(z+\mathrm{tg}^2\frac{k\Pi}{2(n+1)}\right)\right)^{-1}.$$

But, for $\mathcal{R}\mathrm{e}\,(z)\geqq 0$, we have

$$\left|z+\mathrm{tg}^2\frac{k\Pi}{2(n+1)}\right|\geqq \mathrm{tg}^2\frac{k\Pi}{2(n+1)}.$$

Then, for $\mathcal{R}\mathrm{e}\,(z)\geqq 0$

$$|\delta_n(z)|\leqq\left(\prod_{k=1}^{n}|1+e^{ik\pi/(n+1)}|^2\,\mathrm{tg}^2\frac{k\pi}{2(n+1)}\right)^{-1}=\delta_n(0).$$

But, as $\psi(0)=1$, we deduce from (3.4) that

$$\delta_n(0)=2(f_n(0)-f(0))=\frac{1}{n+1}.$$

Consequently,

$$\forall z/\mathcal{R}\mathrm{e}\,(z)\geqq 0 \quad |f_n(z)-f(z)|\leqq\frac{1}{n+1}\left|\frac{(1-\sqrt{z})^{2n}}{(1+\sqrt{z})(1+\psi(z))}\right|.$$

Then, using

$$(1+\sqrt{z})(1+\psi(z))=2\left(\frac{1+\sqrt{z}}{1-\sqrt{z}}\right) \quad \forall|z|\leqq 1/\mathcal{R}\mathrm{e}\,z\geqq 0, \qquad |1-\sqrt{z}|\leqq 1,$$

we easily obtain

$$\forall z/\mathcal{R}\mathrm{e}(z)\geqq 0 \text{ and } |z|\leqq 1\ |f_n(z)-f(z)|\leqq\frac{1}{2(n+1)}\left|\frac{1-\sqrt{z}}{1+\sqrt{z}}\right||1-\sqrt{z}|^{2n}\leqq\frac{1}{2(n+1)}.$$

Using (1), the result follows immediately. □

Now, using Theorems 3.1 and 3.2, and Lemma 3.3, we can summarize our results in Theorem 3.3.

THEOREM 3.3. *Let* $f_n(z)=f^*_{2n+1}(z)$, *where the sequence* f^*_n *is defined by* (3.2); *then we have the following*:

(i) $$f_n(i\omega)=\frac{1}{2(n+1)}+\frac{1}{n+1}\sum_{k=1}^{n}\frac{1}{1+i\omega\,\operatorname{cotg}^2(k\Pi/2(n+1))}.$$

(ii) $$\sup_{\omega\in\mathbb{R}}|f_n(i\omega)-f(i\omega)|=\frac{1}{2(n+1)}.$$

(iii) For $0<\omega_*<\omega^*<+\infty$, there exists a constant $\alpha(\omega_*,\omega_*)\in\,]0,1[$ $/\sup_{[\omega_*,\omega^*]}|f_n(i\omega)-f(i\omega)|\leqq C(\omega_*,\omega^*)\cdot\alpha(\omega_*,\omega^*)^n$.

In Fig. 1 we illustrate the convergence of the sequence $f^*_n(i\omega)$ to $f(i\omega)$, for $\omega\in\mathbb{R}$.

3.3. The error estimate. We now consider the approximate law σ_n defined by

$$\sigma_n(x,\omega)=\sigma_\infty(x)+[\lambda(x)-1]\sigma_0(x)f_n\left(i\frac{\omega}{\omega_c(x)}\right)$$

where f_n is the sequence defined in § 3.2.

From Theorem 3.3 and properties of the functions $\lambda(x)$, $\sigma_0(x)$, and $\omega_c(x)$ we deduce

$$\|\sigma_n-\sigma\|_\infty\leqq\frac{\sigma^*(\lambda^*-1)}{2(n+1)}.$$

This proves that $\Sigma_w(\Omega;\mathbb{R})\subset\overline{\Sigma_a(\Omega;\mathbb{R})}^{\Sigma(\Omega;\mathbb{R})}$. Moreover we have Theorem 3.4.

THEOREM 3.4. *Let* (u,v) *be the solution of problem* (2.1) *associated with Warbourg's law* σ *and let* (u_n,v_n) *be the solution of the same problem* (2.1) *associated with the approximate law* σ_n; *then we have the error estimate*

$$\|(u_n,v_n)-(u,v)\|_{\mathscr{H}(0,T)}\leqq C(\sigma)\frac{\sigma^*(\lambda^*-1)}{2(n+1)}.$$

Comments. The method we have adopted furnishes a constructive way to obtain an approximate polarization law. Of course there is no reason for the law σ_n we consider here to be optimal for a given number n of rational fractions; for example, we can change the initial element of the sequence f^*_n (and take $f^*_0(z)=a\in\mathbb{R}$) without

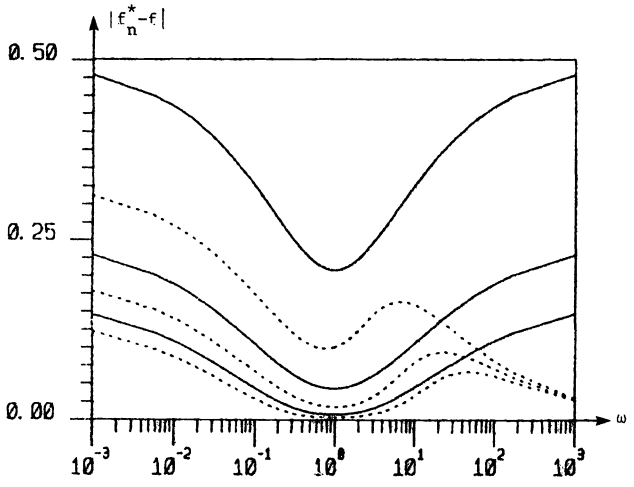


FIG. 1. ——— n odd (1, 3, 5); ---- n even (2, 4, 6).

affecting the rate of convergence of σ_n to σ. In [2], a process for the determination of such an optimal approximate law is described (for any polarization law). It essentially consists of minimizing an appropriate norm of the difference $\sigma_a - \sigma$. Nevertheless, it is clear that the law σ_n we obtain here gives a good initial point for the optimization algorithm. Let us recall that in practice it is sufficient to take $n = 3$, 4 (three simple rational fractions, which implies three auxiliary unknown functions) to get a reasonable approximation of the exact law (see [3]).

4. Conclusion. Our study allowed us to define a good class of polarization laws σ (§ 1) for which the integrodifferential system (0.1) is well posed (§ 2). The approximation result given in § 3 is interesting from both theoretical and practical points of view. It brings a mathematical justification to the numerical method proposed in [3] and gives a way to construct a "good" approximate polarization law.

To conclude, let us mention that looking at result (iii) of Theorem 3.3 we can expect to improve the error estimate given in Theorem 3.4. Moreover our conjecture is that the approximation result we obtain for Warbourg's law could be generalized to any polarization law in the class $\Sigma(\Omega; \mathbb{R})$. A nonconstructive proof using the Stone-Weierstrass Theorem seems to be possible but is still an open and interesting question.

REFERENCES

[1] A. BAMBERGER, B. ENGQUIST, L. HALPERN, AND P. JOLY, *Construction et analyse d'approximations paraxiales en milieu hétérogène, II: Approximation d'ordre supérieur*, Rapport Interne, Ecole Polytechnique, Palaiseau, 1985.

[2] B. COCKBURN, *Etude mathématique et numerique des equations de Maxwell dans des Milieux polarisables*, Thèse de 3ème cycle, Université Paris, Paris, 1983.

[3] ———, *Resolution of Maxwell's equations in polarizable media at radio and lower frequencies*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 843-852.

[4] B. COCKBURN AND P. JOLY, *Justification théorique d'une méthode de résolution des équations de Maxwell en milieu 1D polarisable*, Rapport INRIA 299, Institut National de Recherche en Informatique en Automatique, Le Chesnay, France, 1984.

[5] C. A. DIAS, *Analytical model for a polarisable medium at radio and lower frequencies*, J. Geophysical Res., 77 (1972).

[6] Y. GOLDMAN, *Etude de sensibilité d'un modèle de polarisation électromagnétique vis-à-vis de ses paramètres*, Rapport INRIA 465, Institut National de Recherche en Informatique en Automatique, Le Chesnay, France, 1985.

[7] E. L. LINDMANN, *Free-space boundary conditions for the time-dependent wave equation*, J. Comput. Phys., 18 (1975).

[8] J. L. LIONS AND E. MAGNES, *Problème aux limites non homogènes et applications*, Tome 1, Dunod, Paris 1968.

[9] H. S. WALL, *Analytic Theory of Continued Fractions*, Chelsea, New York, 1948.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

011

FORMATION OF SHOCKS FOR A SINGLE CONSERVATION LAW*

SHIZUO NAKANE†

Abstract. The initial value problem for an equation of scalar conservation law in several space dimensions is considered. By the method of characteristics, the solution of this problem with C^∞-initial datum is concretely constructed. Generally, this solution becomes multivalued in finite time. By virtue of the theory of singularities of C^∞-mappings, its structure as a multivalued function is completely revealed. The entropy solution is constructed by making it single-valued. In this process, shocks occur. Shock surfaces are constructed by using the stable manifold theory. Thus propagation of shocks is described.

Key words. entropy condition, singularities of C^∞-mappings, stable manifold

AMS(MOS) subject classifications. 35L65, 35L67

1. Introduction. In this paper, we will consider the Cauchy problem for a single conservation law:

$$\begin{cases} \dfrac{\partial u}{\partial t}+\sum_{i=1}^{n}\dfrac{\partial f_i(u)}{\partial x_i}=0 \quad \text{in } \{(t,x)\in\mathbb{R}^{n+1};\ t>0\}, \\ u(0,x)=\varphi(x) \quad \text{on } \mathbb{R}^n. \end{cases} \tag{1.1}$$

Here, $f=(f_1,\cdots,f_n)$ is a C^∞-mapping; $\mathbb{R}\to\mathbb{R}^n$ and φ is a real-valued C^∞ rapidly decreasing function on $\mathbb{R}^n$. Generally, we cannot expect the global existence of C^∞-solutions for (1.1). That is, shocks occur (cf. Conway [3]). The purpose of this paper is to investigate concretely how shocks occur and propagate.

Consequently, we consider (1.1) in a weak sense. A weak solution for (1.1) is a function u satisfying the following for any $g\in C_0^\infty(\mathbb{R}^{1+n})$:

$$\iint_{\mathbb{R}^+\times\mathbb{R}^n}\left(u\frac{\partial g}{\partial t}+\sum_{i=1}^{n}f_i(u)\frac{\partial g}{\partial x_i}\right)dt\,dx+\int_{\mathbb{R}^n}\varphi(x)g(0,x)\,dx=0. \tag{1.2}$$

Since uniqueness of weak solutions for (1.1) does not generally hold, we impose an entropy condition in order to eliminate physically meaningless solutions. A weak solution for (1.1) satisfies the entropy condition if the following holds for any $g\in C_0^\infty(\mathbb{R}^{1+n})$, $g\geqq 0$, and for any $k\in\mathbb{R}$,

$$\iint_{\mathbb{R}^+\times\mathbb{R}^n}\operatorname{sgn}(u-k)\left\{(u-k)\frac{\partial g}{\partial t}+\sum_{i=1}^{n}(f_i(u)-f_i(k))\frac{\partial g}{\partial x_i}\right\}dt\,dx\geqq 0. \tag{1.3}$$

We call such a solution an entropy solution. For the existence and uniqueness of entropy solutions, see, for example, Kruzkov [11].

Our main result is that, under a nondegeneracy hypothesis corresponding to that of Guckenheimer [7], we can construct the entropy solution locally and this solution is piecewise C^∞, with C^1-shock surfaces. More precisely, see the theorem in § 6.

Many authors have studied the global structure of shock waves in the case $n=1$ (see, for example, Chen [1], [7], Jennings [10], and Schaeffer [13]). On the other hand, when $n\geqq 2$, there are few results (see Debeneix [4]). These authors either constructed the solutions explicitly by the method of characteristics or they represented the solutions by using the elementary catastrophe theory. Recently Tsuji [14], [15] investigated the

* Received by the editors April 15, 1987; accepted for publication (in revised form) February 17, 1988.
† Tokyo Institute of Polytechnics, 1583 Iiyama, Atsugi City, Kanagawa 243-02, Japan.

formation and propagation of singularities of solutions for Hamilton-Jacobi equations in two space dimensions. He used the theory of singularities of mappings of the plane into the plane obtained by Whitney [17]. Our result is much inspired by his, which suggests that we can treat higher-dimensional cases if we use the theory of singularities of mappings in higher-dimensional spaces. Here we carry out this program.

Our result is an extension of that of § 1 in [7] to n-dimensional cases. The method of proof is also a refinement of the argument in [7], which seems to be incomplete.

An equation of single conservation law in two space dimensions arises in petroleum engineering. See Glimm, Marchesin, and McBryan [5] and Wagner [16].

The plan of this paper is as follows: in § 2, we will construct the solution explicitly by the method of characteristics. As the solution becomes multivalued in finite time, in § 3 and § 4, we will clarify its structure as a multivalued function by virtue of the theory of singularities of mappings and make it single valued. Shocks appear in this process. In §§ 5 and 6, we will construct shock surfaces by using the stable manifold theory, which we have surveyed in Appendix A1. In § 7, we will give an example which arises in petroleum engineering. In Appendix A2, we will prove the smoothness of the stable manifolds with respect to the parameter according to the argument in [2].

2. The method of characteristics. We will construct the solution for (1.1) by the method of characteristics. Set $a_i(u)=f_i'(u)$, $1\leqq i\leqq n$. Then the characteristic line associated with (1.1) through $(0, y)$ is the solution line $(t, x(t))$ of the following:

$$dx_i(t)/dt = a_i(u(t, x(t))), \qquad x_i(0)=y_i \quad (1\leqq i\leqq n). \tag{2.1}$$

Since $du(t, x(t))/dt\equiv 0$ if u is a smooth solution for (1.1), $u(t, x(t))\equiv u(0, x(0))=\varphi(y)$. Therefore the solution of (2.1) can be expressed by $x(t)=y+ta(\varphi(y))$.

We define C^∞-mappings Φ_t and Φ associated with these characteristics as follows:

$$\Phi_t:\mathbb{R}^n_y\to\mathbb{R}^n_x, \qquad \Phi_t(y)=y+ta(\varphi(y)),$$

$$\Phi:\mathbb{R}^{1+n}_{(t,y)}\to\mathbb{R}^{1+n}_{(t,x)}, \qquad \Phi(t, y)=(t, \Phi_t(y)).$$

From the classical theory of characteristics, it follows that the solution of (1.1) is expressed by $u(t, x)=\varphi(\Phi_t^{-1}(x))$ and is C^∞ at (t, x) which is not the critical value of φ. Then our next problem is to study the singularities of mapping Φ_t or Φ. We shall consider mapping Φ, which can be treated more easily than Φ_t.

Set $\lambda(y)=\sum_{i=1}^n a_i'(\varphi(y))\partial\varphi/\partial y_i$. Then, a direct calculation shows that the Jacobian $|J(\Phi)|=\det J(\Phi)$ of Φ is equal to $1+t\lambda(y)$ (see [3]). Here $J(\Phi)$ is the Jacobian matrix of Φ. If $\lambda(y)\geqq 0$, there exists a global C^∞-solution since $|J(\Phi)|\geqq 1$. Thus we assume

$$\min_y \lambda(y)=\lambda(y^0)=-M<0. \tag{A.1}$$

Note that, since φ is rapidly decreasing, λ must attain its minimum at some finite point y^0 unless $\lambda(y)\geqq 0$. Set $t^0=1/M$ and $x^0=\Phi_{t^0}(y^0)$. For $t<t^0$, $|J(\Phi)|>0$ and u is C^∞. Therefore, from now on we will consider for $t\geqq t^0$ near (t^0, y^0). The following assumption corresponds to the first nondegeneracy condition of [7]:

(A.2) The singularities of λ are nondegenerate, i.e., $\nabla\lambda(y)=0$ implies rank Hess $\lambda(y)=n$.

From (A.1) and (A.2), we can easily see that Hess $\lambda(y^0)$ is positive definite. We also remark that the same argument follows for y^0: minimal point of $\lambda(y)$.

3. Structure of singularities of mapping Φ. Under assumptions (A.1) and (A.2), we have the following results.

LEMMA 1. *By affine transformations of coordinates, we may assume the following*:

$$a_i'(\varphi(y^0)) = 0 \quad (1 \leqq i \leqq n-1), \tag{3.1}$$

$$a_n'(\varphi(y^0)) > 0, \tag{3.2}$$

$$a_i(\varphi(y^0)) = 0 \quad (1 \leqq i \leqq n-1). \tag{3.3}$$

Proof. Let $a_n^i = a_i'(\varphi(y^0))$ and $a_n = (a_n^1, \cdots, a_n^n)$. Then, from (A.1), $a_n \neq 0$. We choose b_i $(1 \leqq i \leqq n)$ so that $b_n = a_n/|a_n|$ and $b_1, \cdots, b_n$ form an orthonormal basis of $\mathbb{R}^n$. First we consider the transformation

$$T = t,$$

$$X_j = y_j^0 + \sum_{i=1}^{n} b_j^i (x_i - y_i^0) \quad (1 \leqq j \leqq n).$$

Then, we have

$$\begin{aligned} 0 &= \partial u/\partial t + \sum_{i=1}^{n} a_i(u)\partial u/\partial x_i \\ &= \partial u/\partial T + \sum_{j=1}^{n} \left(\sum_{i=1}^{n} b_j^i a_i(u) \right) \partial u/\partial X_j \\ &= \partial u/\partial T + \sum_{j=1}^{n} b_j(u)\partial u/\partial X_j \end{aligned}$$

and $b_j'(\varphi(y^0)) = \sum_{i=1}^{n} b_j^i a_i'(\varphi(y^0)) = 0$ $(1 \leqq j \leqq n-1)$, > 0 $(j = n)$. Thus we obtain (3.1) and (3.2). Next, by the following transformation

$$T = t,$$

$$X_j = x_j - a_j(\varphi(y^0))t \quad (1 \leqq j \leqq n-1),$$

$$X_n = x_n,$$

(3.3) is easily seen to be satisfied. This completes the proof. □

Remark 1. It is easy to see that the above transformations preserve assumptions (A.1) and (A.2).

Before stating the proposition, we need a definition.

DEFINITION 1. The singularity y^0 of a C^∞-mapping $M: \mathbb{R}_y^m \to \mathbb{R}_x^m$ is said to be fold (respectively, cusp), if, by diffeomorphisms in $\mathbb{R}_y^m$ and in $\mathbb{R}_x^m$ sending y^0 and $M(y^0)$ to the origins, M is transformed into the following form:

$$\begin{array}{l} x_1 = y_1, \\ \vdots \\ x_{m-1} = y_{m-1}, \\ x_m = y_m^2 \end{array} \quad \left(\text{respectively,} \quad \begin{array}{l} x_1 = y_1, \\ \vdots \\ x_{m-1} = y_{m-1}, \\ x_m = y_m^3 - y_1 y_m. \end{array} \right)$$

PROPOSITION 2. *Near $t = t^0$, the singularities of Φ must be fold or cusp.*

Proof. Consider the C^∞-mapping $h: \mathbb{R}_{(t,y)}^{1+n} \to \mathbb{R}_{(T,Y)}^{1+n}$,

$$T = t,$$

$$Y_i = y_i + t a_i(\varphi(y)) \quad (1 \leqq i \leqq n-1),$$

$$Y_n = y_n.$$

From (3.1), it follows that h is a diffeomorphism near (t^0, y^0). We represent h^{-1} by

$$t = T,$$
$$y_i = b_i(T, Y) \qquad (1 \leqq i \leqq n),$$

where $b_n(T, Y) = Y_n$ and b_j satisfy

$$Y_j = b_j(T, Y) + Ta_j(\varphi(b(T, Y))) \qquad (1 \leqq j \leqq n-1) \tag{3.4}$$

Then, $\tilde{\Phi} = \Phi \circ h^{-1}$ is expressed by

$$t = T,$$
$$x_j = Y_j \qquad (1 \leqq j \leqq n-1),$$
$$x_n = Y_n + Ta_n(\varphi(b(T, Y))).$$

Note that (T, Y) is the "adapted system of coordinates" in the terminology of Morin [12]. We only have to show that the singularities of $\tilde{\Phi}$ must be cusp unless they are fold. In order to do so, we use the characterization of fold points and cusp points obtained in [12]. That is, we have only to show

$$\partial^3 x_n / \partial Y_n^3 \neq 0, \tag{3.5}$$

$$\partial^2 x_n / \partial T \partial Y_n \neq 0, \tag{3.6}$$

on the set

$$\Sigma^{1,1} = \{(T, Y);\ \partial x_n / \partial Y_n = \partial^2 x_n / \partial Y_n^2 = 0\}.$$

Since

$$\partial x_n / \partial Y_n = |J(\tilde{\Phi})| = |J(\Phi)| / |J(h)| = \{1 + T\lambda(b)\} / |J(h)|,$$
$$\Sigma^{1,1} = \{(T, Y);\ 1 + T\lambda(b) = \textstyle\sum_{j=1}^{n} \partial\lambda / \partial y_j \cdot \partial b_j / \partial Y_n = 0\}.$$

On this set, we have

$$\partial^3 x_n / \partial Y_n^3 = T\left(t_b \cdot \mathrm{Hess}\, \lambda \cdot b + \sum_{i=1}^{n} \partial\lambda / \partial y_i \cdot \partial b_i / \partial Y_n \right) / |J(h)|,$$

$$\partial^2 x_n / \partial T \partial Y_n = \left(\lambda + T \sum_{i=1}^{n} \partial\lambda / \partial y_i \cdot \partial b_i / \partial Y_n \right) \Big/ |J(h)|,$$

where $b = {}^t(\partial b_1 / \partial Y_n, \cdots, \partial b_n / \partial Y_n)$. From (A.1), (A.2), (3.1), and (3.2), we have $\partial^3 x_n / \partial Y_n^3 > 0$ and $\partial^2 x_n / \partial T \partial Y_n < 0$ on $\Sigma^{1,1}$. This completes the proof. □

Remark 2. The following facts, which will be used later, can be easily seen:

$$\partial^2 x_n / \partial Y_j \partial Y_n = 0 \quad \text{at } (T^0, Y^0) = h(t^0, y^0) \quad \text{for any } j. \tag{3.7}$$

These imply that the singularity y^0 of mapping Φ_{t^0} is neither fold nor cusp. Thus we consider Φ instead of Φ_t. On the other hand, it is shown by the same argument as in Proposition 2 that, for $t > t^0$, the singularities of Φ_t must be fold or cusp.

Remark 3. When $n = 1$, Chen [1] has proved the same result.

Remark 4. Tsuji [14], [15] studied the singularities of mapping H_t, which corresponds to Φ_t, for the Hamilton-Jacobi equation. He assumed that the singularities of H_t must be fold or cusp.

We set $\Sigma^1 = \{(T, Y) \in \mathbb{R}^{n+1};\ 1 + T\lambda(b(T, Y) = 0\}$. Then, Σ^1 (respectively, $\Sigma^{1,1}$) is the set of singularities (respectively, cusp points) of $\tilde{\Phi}$.

LEMMA 3. *Near* (T^0, Y^0), $\Sigma^{1,1}$ *is a* C^∞*-submanifold of* $\mathbb{R}^{n+1}$ *with codimension* 2 *parametrized by* $Y' = (Y_1, \cdots, Y_{n-1})$.

This lemma follows from the implicit function theorem. The following lemma is also easy to see.

LEMMA 4. $\tilde{\Phi}(\Sigma^{1,1})$ *is a* C^∞*-submanifold of* $\mathbb{R}^{n+1}$ *with codimension* 2 *parametrized by* x'.

We represent $\Sigma^{1,1} = \{(T, Y);\ T = \tilde{\alpha}(Y'),\quad Y_n = \tilde{\beta}(Y')\}$ and $\tilde{\Phi}(\Sigma^{1,1}) = \{(t, x);\ t = \alpha(x'),\ x_n = \beta(x')\}$.

4. Formation of shocks. From the argument of the preceding section, the structure of singularities of mapping $\tilde{\Phi}$ (and therefore of Φ) has been completely clarified. From the canonical form of mappings at cusp points (cf. Definition 1), it follows that the inverses $\Phi^{-1}, \tilde{\Phi}^{-1}$ are triple-valued. See Fig. 1.

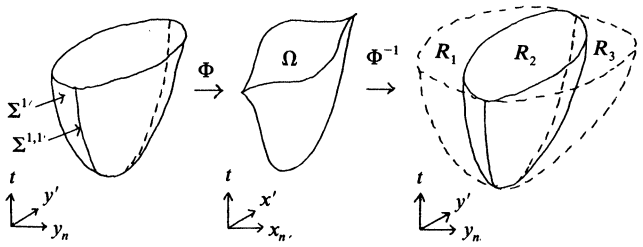


FIG. 1

We set $\Omega = \tilde{\Phi}(\{1 + T\lambda(b(T, Y)) < 0\})$, and we write the three inverse images of Ω by $\tilde{R}_i$ $(1 \leqq i \leqq 3)$ and three branches of $\tilde{\Phi}^{-1}$ in Ω by

$$(\tilde{\Phi}|_{\tilde{R}_i})^{-1}: \begin{cases} T = t, \\ Y' = x', \\ Y_n = \tilde{g}_n^{(i)}(t, x), \end{cases}$$

where $\tilde{g}_n^{(1)}(t, x) < \tilde{g}_n^{(2)}(t, x) < \tilde{g}_n^{(3)}(t, x)$ in Ω. Since $\Phi = \tilde{\Phi} \circ h$, we can write three branches of Φ^{-1} by

$$t = t, \qquad y = g^{(i)}(t, x).$$

We set $u_i(t, x) = \varphi(g^{(i)}(t, x))$ in Ω $(1 \leqq i \leqq 3)$. Then we have the following lemma.

LEMMA 5. $\partial u_1/\partial x_n,\ \partial u_3/\partial x_n < 0$, *and* $\partial u_2/\partial x_n > 0$ *in* Ω.

Proof. Set $\tilde{\varphi}(T, Y) = \varphi \circ h^{-1}(T, Y) = \varphi(b(T, Y))$. Then, $u_i(t, x) = \tilde{\varphi}(t, x', g_n^{(i)}(t, x))$ and $\partial u_i/\partial x_n = (\sum_{i=1}^n \partial\varphi/\partial y_i \cdot \partial b_i/\partial Y_n)\partial g_n^{(i)}/\partial x_n$. From (3.1), (3.2), (3.4), and (A.1), it follows that, at (T^0, Y^0), $\partial b_i/\partial Y_n = 0$ $(1 \leqq i \leqq n-1)$, $\equiv 1$ $(i = n)$ and at (t^0, y^0), $\partial\varphi/\partial y_n < 0$. Then $\partial g_n^{(i)}/\partial x_n = \partial Y_n/\partial x_n = (\partial x_n/\partial Y_n)^{-1} > 0$ $(i = 1, 3)$, < 0 $(i = 2)$. Thus we get $\partial u_i/\partial x_n < 0$ $(i = 1, 3)$, > 0 $(i = 2)$ in Ω near (T^0, Y^0). This completes the proof.

Thus we have completely clarified the structure of the solution u of (1.1) as a multivalued function.

Since we are looking for a single-valued solution, we must make u single valued in Ω. For this reason, u must be discontinuous in Ω. We consider a weak solution having C^1-shock surface $S = \{\nu(t, x) = 0\}$ near (t^0, x^0). Let $\mathbf{n} = \mathbf{n}(p)$ be the unit normal vector to S at $p \in S$ and let $u_\pm(p) = \lim_{\varepsilon \to +0} u(p \pm \varepsilon\mathbf{n})$. Then the Rankine-Hugoniot condition (4.1) follows from (1.2):

$$\mathbf{n} \cdot F(u_+, u_-) = 0, \tag{4.1}$$

where $F(u_+, u_-) = ([u], [f_1(u)], \cdots, [f_n(u)])$, $[u] = u_+ - u_-$, $[f_i(u)] = f_i(u_+) - f_i(u_-)$. If $S = \{\nu(t, x) = 0\}$, (4.1) is expressed by

$$\frac{\partial \nu}{\partial t} + \sum_{i=1}^{n} \frac{[f_i(u)]}{[u]} \cdot \frac{\partial \nu}{\partial x_i} = 0 \quad \text{on } S. \tag{4.2}$$

Furthermore, since S issues from $\tilde{\Phi}(\Sigma^{1,1})$, ν must satisfy

$$\nu(\alpha(x'), x', \beta(x')) = 0. \tag{4.3}$$

On the other hand, if we orient $\mathbf{n}$ so that $u_+ > u_-$, the entropy condition implies

$$\mathbf{n} \cdot F(k, u_+) \geqq 0, \tag{4.4}$$

for any k, $u_- \leqq k \leqq u_+$.

Later we will construct S so that S is parametrized by (t, x'), i.e., S is expressed by $x_n = \psi(t, x')$. Then, (4.3)-(4.4) implies

$$\frac{\partial \psi}{\partial t} + \sum_{i=1}^{n-1} \frac{[f_i(u)]}{[u]} \frac{\partial \psi}{\partial x_i} = \frac{[f_n(u)]}{[u]} \quad \text{on } S, \tag{4.5}$$

$$\psi(\alpha(x'), x') = \beta(x'). \tag{4.6}$$

Now we define, in Ω, a single-valued function $u(t, x)$ by

$$u(t, x) = \begin{cases} u_1(t, x) & \text{if } x_n < \psi(t, x'), \\ u_3(t, x) & \text{if } x_n > \psi(t, x'). \end{cases} \tag{4.7}$$

And let $\psi = \psi(t, x')$ be the solution of (4.5), (4.6) with $u_+ = u_1$ and $u_- = u_3$. Then we have the following lemma.

LEMMA 6. *Suppose $\partial\psi/\partial x_i$ $(1 \leqq i \leqq n-1)$ are bounded in Ω. Then the entropy condition is satisfied.*

Proof. From the definition, $\mathbf{n} = (\partial\psi/\partial t, \partial\psi/\partial x', -1)$. Hence (4.4) is equivalent to

$$\frac{\partial \psi}{\partial t} + \sum_{i=1}^{n-1} \frac{f_i(k) - f_i(u_1)}{k - u_1} \frac{\partial \psi}{\partial x_i} \leqq \frac{f_n(k) - f_n(u_1)}{k - u_1} \quad (u_3 \leqq k \leqq u_1), \tag{4.8}$$

where $u_i = u_i(t, x', \psi(t, x'))$.

It follows from (4.5) that (4.8) is equivalent to

$$\begin{aligned} &\sum_{i=1}^{n-1} \left(\frac{f_i(u_1) - f_i(k)}{u_1 - k} - \frac{f_i(u_1) - f_i(u_3)}{u_1 - u_3} \right) \frac{\partial \psi}{\partial x_i} \\ &\leqq \frac{f_n(u_1) - f_n(k)}{u_1 - k} - \frac{f_n(u_1) - f_n(u_3)}{u_1 - u_3} \quad (u_3 \leqq k \leqq u_1). \end{aligned} \tag{4.9}$$

From (3.2) and the assumption, $|\partial\psi/\partial x_i| \leqq C$ $(1 \leqq i \leqq n-1)$ and $a_n'(u) \geqq c$ in some small neighborhood of (t^0, x^0) for some C, $c > 0$. And it follows from (3.2) that, for any $\varepsilon > 0$, there exists an open neighborhood of (t^0, x^0) such that we have $|a_i'(u)| \leqq \varepsilon$ there. Then the following lemma shows:

$$|\text{The left-hand side of (4.9)}| \leqq \frac{1}{2} \varepsilon C (k - u_3),$$

$$\text{The right-hand side of (4.9)} \geqq \frac{c}{2} (k - u_3),$$

which is easily seen. By taking ε sufficiently small, we have (4.9). This completes the proof. □

LEMMA 7. *Let* $g(u)=\{f(u)-f(w)\}/(u-w)$. *Then we have the following*:
(i) *if* $f''(u)\geqq c>0$, $g'(u)\geqq c/2$,
(ii) *if* $|f''(u)|\leqq\varepsilon$, $|g'(u)|\leqq\varepsilon/2$.

Later we shall prove the existence of ψ and the boundedness of $\partial\psi/\partial x_i$.

Remark 5. When Tsuji [14], [15] made the solution u of Hamilton-Jacobi equation single valued, he proved that u must jump from u_1 to u_3.

5. Construction of shock surfaces (I). Now we will construct the soution of (4.2)-(4.3) or (4.5)-(4.6). It is a noncharacteristic Cauchy problem of a first-order nonlinear partial differential equation. Note that, since $\tilde{\Phi}(\Sigma^1)$ is the set of critical values of $\tilde{\Phi}$, u_i are not Lipschitz continuous on $\tilde{\Phi}(\Sigma^1)$. Thus we cannot use the usual method of characteristics. According to the argument in [7], we pull back the problem by $\tilde{\Phi}$. We define a vector field X in Ω by

$$X=(1+t\lambda(y))\left(\frac{\partial}{\partial t}+\sum_{i=1}^{n}\frac{[f_i(u)]}{[u]}\frac{\partial}{\partial x_i}\right), \tag{5.1}$$

where $[u]=u_1(t,x)-u_3(t,x)$, etc. and a vector field $\tilde{X}$ in $\tilde{R}_2$ by

$$\begin{aligned}\tilde{X}=&(1+T\lambda(b(T,Y)))\frac{\partial}{\partial T}+(1+T\lambda)\sum_{i=1}^{n-1}\frac{[f_i(u)]}{[u]}\frac{\partial}{\partial Y_i}\\&+|J(h)|^{-1}\left\{(1+T\lambda)\left(\frac{[f_n(u)]}{[u]}-a_n(\varphi)\right)\right.\\&\left.-Ta_n'(\varphi)\sum_{k=1}^{n}\frac{\partial\varphi}{\partial y_k}\left(\frac{[f_k(u)]}{[u]}-a_k(\varphi)\right)\right\}\frac{\partial}{\partial Y_n},\end{aligned} \tag{5.2}$$

where $[u]=u^{(1)}-u^{(3)}$, etc., $u^{(i)}=u^{(i)}(T,Y)=\tilde{\varphi}\circ(\tilde{\Phi}|_{\tilde{R}_i})^{-1}\circ\tilde{\Phi}(T,Y)$ $(i=1,3)$. By a direct calculation, we have $X=\tilde{\Phi}_*\tilde{X}$. In this section, we will investigate some properties of $\tilde{X}$.

LEMMA 8. $\tilde{X}$ *can be continued, as a* C^∞*-vector field, to a neighborhood of* (T^0, Y^0).

Proof. Since the smoothness of a vector field is preserved under diffeomorphisms, we may assume that $\tilde{\Phi}$ is the canonical form at cusp points. That is, we may take $\tilde{\Phi}$ as follows:

$$t=T,\quad x'=Y',\quad x_n=Y_n^3-TY_n.$$

What we have to show is the smoothness of $[f_i(u)]/[u]$. Since the roots of $\tilde{Y}_n^3-T\tilde{Y}_n=Y_n^3-TY_n$ are $\tilde{Y}_n=Y_n$, $(-Y_n\pm\Delta)/2$, where $\Delta=(4T-3Y_n^2)^{1/2}$, we have

$$u^{(1)}(T,Y)=\tilde{\varphi}\left(T,Y',\frac{-Y_n-\Delta}{2}\right),\qquad u^{(3)}(T,Y)=\tilde{\varphi}\left(T,Y',\frac{-Y_n+\Delta}{2}\right).$$

Then

$$\begin{aligned}\frac{[f_i]}{[u]}&=\frac{1}{u^{(1)}-u^{(3)}}\int_0^1\frac{d}{ds}f_i(su^{(1)}+(1-s)u^{(3)})\,ds\\&=\int_0^1 a_i\left(s\tilde{\varphi}\left(T,Y',\frac{-Y_n-\Delta}{2}\right)+(1-s)\tilde{\varphi}\left(T,Y',\frac{-Y_n+\Delta}{2}\right)\right)ds\\&=F_i(T,Y,\Delta).\end{aligned}$$

From the hypothesis, F_i is C^∞ with respect to (T,Y,Δ) and $F_i(T,Y,-\Delta)=F_i(T,Y,\Delta)$. Then the function $G_i(T,Y,\xi)=F_i(T,Y,\xi^{1/2})$, defined in $\xi>0$, is smoothly continued

to $\xi \leqq 0$. Now we have $F_i(T, Y, \Delta) = G_i(T, Y, \Delta^2) = G_i(T, Y, 4T - 3Y_n^2)$, which implies $[f_i]/[u]$ is smoothly extended to a neighborhood of (T^0, Y^0). This completes the proof. □

LEMMA 9. *Let Σ be the set of singularities of the vector field $\tilde{X}$. Then $\Sigma = \Sigma^{1,1}$.*

Proof. It is easy to see $\Sigma^{1,1} \subset \Sigma \subset \Sigma^1$. Hence we show $\Sigma \subset \Sigma^{1,1}$. Note that $a_i(u) = a_i(\tilde{\varphi}) + a_i'(\tilde{\varphi})(u - \tilde{\varphi}) + (u - \tilde{\varphi})^2 \int_0^1 (1-s) a_i''(\tilde{\varphi} + s(u - \tilde{\varphi}))\, ds$. We set

$$\begin{aligned}
\frac{[f_i]}{[u]} - a_i(\tilde{\varphi}) &= \frac{1}{u^{(1)} - u^{(3)}} \int_{u^{(3)}}^{u^{(1)}} \{a_i(u) - a_i(\tilde{\varphi})\}\, du \\
&= \frac{a_i'(\tilde{\varphi})}{u^{(1)} - u^{(3)}} \int_{u^{(3)}}^{u^{(1)}} (u - \tilde{\varphi})\, du \\
&\quad + \frac{1}{u^{(1)} - u^{(3)}} \int_{u^{(3)}}^{u^{(1)}} (u - \tilde{\varphi})^2 \int_0^1 (1-s) a_i''(\tilde{\varphi} + s(u - \tilde{\varphi}))\, ds\, du \\
&= I_1 + I_2.
\end{aligned}$$

Then

$$I_1 = \frac{1}{2} a_i'(\tilde{\varphi})(u^{(1)} + u^{(3)} - 2\tilde{\varphi}),$$

$$\begin{aligned}
|I_2| &\leqq \frac{C}{u^{(1)} - u^{(3)}} \int_{u^{(3)}}^{u^{(1)}} (u - \tilde{\varphi})^2\, du \\
&\leqq C' |(u^{(1)} - \tilde{\varphi}, u^{(3)} - \tilde{\varphi})|^2.
\end{aligned}$$

Thus we have

$$\begin{aligned}
&- T a_n'(\tilde{\varphi}) \sum_{k=1}^{n} \frac{\partial \varphi}{\partial y_k} \left(\frac{[f_k]}{[u]} - a_k \right) \\
&\quad = - T a_n'(\tilde{\varphi}) \sum_{k=1}^{n} \frac{\partial \varphi}{\partial y_k} \left\{ \frac{1}{2} a_k'(\tilde{\varphi})(u^{(1)} + u^{(3)} - 2\tilde{\varphi}) + O(|(u^{(1)} - \tilde{\varphi}, u^{(3)} - \tilde{\varphi})|^2) \right\} \\
&\quad = - \frac{T}{2} a_n'(\tilde{\varphi}) \{ \lambda (u^{(1)} + u^{(3)} - 2\tilde{\varphi}) + O(|(u^{(1)} - \tilde{\varphi}, u^{(3)} - \tilde{\varphi})|^2) \}.
\end{aligned}$$

Since $u^{(1)} = \tilde{\varphi} \neq u^{(3)}$ or $u^{(1)} \neq \tilde{\varphi} = u^{(3)}$ on $\Sigma^1 - \Sigma^{1,1}$, $|u^{(1)} + u^{(3)} - 2\tilde{\varphi}| = |(u^{(1)} - \tilde{\varphi}, u^{(3)} - \tilde{\varphi})|$ on $\Sigma^1 - \Sigma^{1,1}$. These imply that the points in $\Sigma^1 - \Sigma^{1,1}$ are not singularities of $\tilde{X}$. This completes the proof.

Next we shall compute the Jacobian matrix $J(\tilde{X})$ of $\tilde{X}$ at (T^0, Y^0). To do so, we study how it is exchanged by diffeomorphisms. Suppose that a vector field $X_1 = \sum c_i(x)(\partial/\partial x_i)$ in $\mathbb{R}^n$, is, by a diffeomorphism: $x = g(y)$, transformed into $X_2 = \sum d_j(y)(\partial/\partial y_j)$. Then we have the following lemma

LEMMA 10. *At singularities of X_1, $J(X_2) = J(g)^{-1} J(X_1) J(g)$.*

We omit the proof. Now we will transform $\tilde{\Phi}$ into the canonical form at cusp points by diffeomorphisms. We set $g(T, Y) = Y_n + T a_n(b(T, Y))$. Then from the proof of Proposition 2, we have $\partial g/\partial Y_n = \partial^2 g/\partial Y_n^2 = 0$ and $\partial^3 g/\partial Y_n^3$, $\partial^2 g/\partial T \partial Y_n \neq 0$ on $\Sigma^{1,1}$. The unfolding theorem (see [13]) shows that there exist C^∞-functions $h(T, Y)$ near (T^0, Y^0), $a_0(T, Y')$, $a_1(T, Y')$ near $(T^0, Y^{0\prime})$ such that

$$g(T, Y) = h(T, Y)^3 - a_1(T, Y') h(T, Y) + a_0(T, Y')$$

and $\partial h/\partial Y_n \neq 0$. By a direct calculation, we have $\Sigma^{1,1} = \{(T, Y);\ h(T, Y) = a_1(T, Y') = 0\}$ and $\partial a_1/\partial T \neq 0$ on $\Sigma^{1,1}$.

Consider the following mapping $P:\mathbb{R}^{n+1}\to\mathbb{R}^{n+1}$,

$$s=a_1(T, Y'),\quad p'=Y',\quad p_n=h(T, Y).$$

Then, $J(P)$ is equal to

$$\begin{bmatrix}\partial a_1/\partial T & \partial a_1/\partial Y' & 0\\ 0 & I_{n-1} & 0\\ \partial h/\partial T & \partial h/\partial Y' & \partial h/\partial Y_n\end{bmatrix}.$$

From (3.7) it follows that $\partial a_1/\partial Y'=0$ at (T^0, Y^0), which implies that $J(P)(T^0, Y^0)$ is equal to

$$\begin{bmatrix}\partial a_1/\partial T & 0 & 0\\ 0 & I_{n-1} & 0\\ \partial h/\partial T & \partial h/\partial Y' & \partial h/\partial Y_n\end{bmatrix}$$

and $|J(P)|(T^0, Y^0)=\partial a_1/\partial T\cdot\partial h/\partial Y_n\neq 0$. Hence P is a diffeomorphism near (T^0, Y^0). We write its inverse P^{-1} by

$$T=T(s, p'),\quad Y'=p',\quad Y_n=Y_n(s, p).$$

Then, $\tilde{\Phi}'=\tilde{\Phi}\circ P^{-1}$ is expressed by

$$t=T(s, p'),\quad x'=p',\quad x_n=p_n^3-sp_n+a_0(T(s, p'), p').$$

Now we set a C^∞-mapping $Q:\mathbb{R}^{n+1}_{(t,x)}\to\mathbb{R}^{n+1}_{(r,q)}$:

$$r=a_1(t, x'),\quad q'=x',\quad q_n=x_n-a_0(t, x').$$

Since $J(Q)(t^0, x^0)$ is equal to

$$\begin{bmatrix}\partial a_1/\partial t & 0 & 0\\ 0 & I_{n-1} & 0\\ * & * & 1\end{bmatrix},$$

Q is a diffeomorphism near (t^0, x^0). Then the C^∞-mapping $\tilde{\Phi}''=Q\circ\tilde{\Phi}'$ is expressed by

$$r=s,\quad q'=p',\quad q_n=p_n^3-sp_n,$$

which is the desired canonical form.

Now we write $\tilde{X}=c_0(T, Y)(\partial/\partial T)+\sum_{i=1}^n c_i(T, Y)(\partial/\partial Y_i)$ and $\tilde{X}'=P_*\tilde{X}=d_0(s, p)(\partial/\partial s)+\sum_{i=1}^n d_i(s, p)(\partial/\partial p_i)$.

LEMMA 11. *At* $(s^0, p^0)=P(T^0, Y^0)$, $J(\tilde{X}')$ *is equal to*

$$\begin{bmatrix}\lambda(y^0) & 0 & 0\\ 0 & 0_{n-1} & 0\\ * & * & -\frac{3}{2}\lambda(y^0)\end{bmatrix}.$$

Proof. Since c_i $(0\leqq i\leqq n)$, $\partial a_1/\partial Y_i$ $(1\leqq i\leqq n-1)$, $\partial c_0/\partial Y_n=0$ at (T^0, Y^0) and $d_0=c_0(\partial a_1/\partial T)+\sum_{i=1}^{n-1} c_i(\partial a_1/\partial Y_i)$, we have $(\partial d_0/\partial s)=(\partial a_1/\partial T)\cdot(\partial c_0/\partial T)\cdot(\partial T/\partial s)=\lambda$ at (s^0, p^0). In the same way, we get $\partial d_0/\partial p_j=0$ at (s^0, p^0) $(1\leqq j\leqq n)$.

Next note that $d_i=c_i=(1+T\lambda)([f_i(u)]/[u])$ $(1\leqq i\leqq n-1)$ and $[f_i(u)]/[u]=a_i(\tilde{\varphi})$ on $\Sigma^{1,1}$. Then, from (3.3), $\partial d_i/\partial s=\partial d_i/\partial p_j=0$ at (s^0, p^0) $(1\leqq i\leqq n-1,\ 1\leqq j\leqq n)$.

Now we consider $d_n = c_0(\partial h/\partial T) + \sum_{j=1}^{n} c(\partial h/\partial Y_j)$. By the same argument as above, we have only to consider the term $c_n(\partial h/\partial Y_n)$. If we set $\tilde{\tilde{\varphi}} = \tilde{\varphi} \circ P^{-1}$, $u^{(1)} = u^{(1)}(s, p) = \tilde{\tilde{\varphi}}(s, p', (-p_n - \Delta)/2)$, $u^{(3)} = u^{(3)}(s, p) = \tilde{\tilde{\varphi}}(s, p', (-p_n + \Delta)/2)$, where $\Delta = (4s - 3p_n^2)^{1/2}$. The same argument as in the proof of Lemma 8 shows

$$\begin{aligned}\frac{[f_i]}{[u]} &= \int_0^1 a_i\left(r\tilde{\tilde{\varphi}}\left(s, p', \frac{-p_n - \Delta}{2}\right) + (1-r)\tilde{\tilde{\varphi}}\left(s, p', \frac{-p_n + \Delta}{2}\right)\right) dr \\ &= \tilde{F}_i(s, p, \Delta) \\ &= \tilde{G}_i(s, p, 4s - 3p_n^2),\end{aligned}$$

where $\tilde{F}_i(s, p, \Delta)$, $\tilde{G}_i(s, p, \xi)$ are C^∞. Then, on $P(\Sigma^{1,1}) = \{s = p_n = 0\}$,

$$\frac{\partial}{\partial p_n}\frac{[f_i]}{[u]} = \frac{\partial \tilde{F}_i}{\partial p_n} - 6p_n\frac{\partial \tilde{G}_i}{\partial p_n} = -\frac{1}{2}a_i'(\tilde{\tilde{\varphi}}(0, p', 0))\frac{\partial \tilde{\tilde{\varphi}}}{\partial p_n}(0, p', 0).$$

Hence, at (s^0, p^0),

$$\frac{\partial}{\partial p_n}\left\{\frac{[f_i]}{[u]} - a_i\right\} = \begin{cases} 0 & (1 \leqq i \leqq n-1), \\ -\dfrac{3}{2}a_n'(\tilde{\tilde{\varphi}}(0, p', 0))\dfrac{\partial \tilde{\tilde{\varphi}}}{\partial p_n}(0, p', 0) & (i = n). \end{cases}$$

Here we have

$$\frac{\partial \tilde{\tilde{\varphi}}}{\partial p_n} = \frac{\partial \tilde{\varphi}}{\partial Y_n}\frac{\partial Y_n}{\partial p_n} = \frac{\partial \varphi}{\partial y_n}\Big/\frac{\partial h}{\partial Y_n} \qquad \text{at } (s^0, p^0).$$

Consequently,

$$\begin{aligned}\frac{\partial d_n}{\partial p_n} &= \frac{\partial h}{\partial Y_n}|J(h)|^{-1}\left\{-Ta_n'\sum_{j=1}^{n}\frac{\partial \varphi}{\partial y_j}\frac{\partial}{\partial p_n}\left(\frac{[f_j]}{[u]} - a_j\right)\right\} \\ &= -Ta_n'\frac{\partial \varphi}{\partial y_n}\left\{-\frac{3}{2}a_n'(\varphi(y^0))\frac{\partial \varphi}{\partial y_n}(y^0)\right\} \\ &= -T\lambda(y^0)\left(-\frac{3}{2}\lambda(y^0)\right) \\ &= -\frac{3}{2}\lambda(y^0).\end{aligned}$$

This completes the proof. □

From Lemmas 10 and 11, we have the following lemma.

LEMMA 12. *$J(\tilde{X})(T^0, Y^0)$ is equal to*

$$\begin{bmatrix} \lambda(y^0) & 0 & 0 \\ 0 & 0_{n-1} & 0 \\ * & * & -\frac{3}{2}\lambda(y^0) \end{bmatrix}. \tag{5.3}$$

6. Construction of shock surfaces (II). Recall that we are looking for a solution to (4.2)-(4.3) or (4.5)-(4.6). Instead of (4.2), we consider

$$(1 + T\lambda)\left\{\frac{\partial \nu}{\partial t} + \sum_{i=1}^{n}\frac{[f_i(u)]}{[u]}\cdot\frac{\partial \nu}{\partial x_i}\right\} = 0. \tag{6.1}$$

If $1 + T\lambda \neq 0$ on $\nu = 0$, the solution ν of (6.1) satisfies (4.2). Since the coefficients of the vector field X do not have enough regularity, we consider $\tilde{X}$ instead of X. Lemma

8 assures us that $\tilde{X}$ is a C^∞-vector field near (T^0, Y^0). Lemma 9 tells us that $\tilde{X}$ has singularities on $\Sigma^{1,1}$. Lemma 12 suggests that we can make use of the stable manifold theory in order to obtain integral curves of $\tilde{X}$ tending to $\Sigma^{1,1}$.

Recall that $\tilde{X} = c_0(T, Y)(\partial/\partial T) + \sum_{i=1}^n c_i(T, Y)(\partial/\partial Y_i)$ and $\Sigma^{1,1} = \{(T, Y);\ T = \tilde{\alpha}(Y'),\ Y_n = \tilde{\beta}(Y')\}$. Thus our purpose is to get a function $\Psi(t) = \Psi(t, y') = {}^t(T(t, y'), Y_1(t, y'), \cdots, Y_n(t, y'))$ satisfying

$$\begin{aligned} &\frac{d}{dt}\Psi(t) = c(\Psi(t)), \\ &\lim_{t\to+\infty} \Psi(t, y') = {}^t(\tilde{\alpha}(y'), y', \tilde{\beta}(y')), \end{aligned} \tag{6.2}$$

where $c = c(T, Y) = {}^t(c_0(T, Y), c_1(T, Y), \cdots, c_n(T, Y))$ and $y' \in \mathbb{R}^{n-1}$ parametrizes $\Sigma^{1,1}$. If we define $u = {}^t(u_0, u_1, \cdots, u_n)$ by

$$\begin{aligned} &u_0(t, y') = T(t, y') - \tilde{\alpha}(y'), \\ &u_j(t, y') = Y_j(t, y') - y_j \quad (1 \leqq j \leqq n-1), \\ &u_n(t, y') = Y_n(t, y') - \tilde{\beta}(y'), \end{aligned} \tag{6.3}$$

then u must satisfy

$$\begin{aligned} &\frac{d}{dt}u = A(y')u + f(y', u), \\ &\lim_{t\to+\infty} u = 0. \end{aligned} \tag{6.4}$$

Here $A(y') = J(\tilde{X})|_{\Sigma^{1,1}}$ and $f(y', u)$ is C^∞ near $(y', u) = (Y^{0\prime}, 0)$ and satisfies $f(y', 0) = 0$, $\nabla_u f(y', 0) = 0$.

We set $u(t, y') = e^{-at}v(t, y')$ for some $a > 0$. Then v satisfies

$$\frac{d}{dt}v = (A(y') + aI)v + e^{at}f(y', e^{-at}v). \tag{6.5}$$

Lemma 12 implies that, by taking a appropriately, the real part of one eigenvalue of $A(y') + aI$ is negative and the real parts of the other n eigenvalues are positive for y' near $Y^{0\prime}$. Then there exist one-dimensional stable manifolds. As for the stable manifold theory, see Coddington and Levinson [2, Chap. 13].

We will state it more precisely. First we blockwise diagonalize $A(y') + aI$. Because of the distinctness of its eigenvalues, there exists a matrix-valued C^∞-function $P(y')$ near $Y^{0\prime}$ satisfying

$$(A(y') + aI)P(y') = P(y')D(y'), \tag{6.6}$$

where

$$D(y') = \begin{bmatrix} d(y') & 0 \\ 0 & D_1(y') \end{bmatrix} = \begin{bmatrix} d(y') & 0 & 0 \\ 0 & D_2(y') & 0 \\ 0 & 0 & e(y') \end{bmatrix}, \tag{6.7}$$

where $D_1(y')$ (respectively, $D_2(y')$) is an $n \times n$ (respectively, $(n-1) \times (n-1)$) matrix (see, for example, Hsieh and Sibuya [9]). Set $P(y') = [p_{ij}(y')]_{0 \leqq i,j \leqq n}$. Then, by a direct calculation, we have

$$p_{00}(Y^{0\prime}) \neq 0 \quad \text{and} \quad p_{i0}(Y^{0\prime}) = 0 \quad (1 \leqq i \leqq n-1). \tag{6.8}$$

If we set $v = P(y')w$, w satisfies

$$\frac{d}{dt}w = D(y')w + g(t, y', w), \tag{6.9}$$

where $g(t, y', w) = P(y')^{-1} e^{at} f(y', e^{-at}P(y')w)$. It is easy to see that g is C^∞ and $g(t, y', 0) = 0$, $\nabla_w g(t, y', 0) = 0$. Now we define $U(t) = U(t, y') = U_0(t, y') + U_1(t, y')$, where

$$U_0(t) = \begin{bmatrix} e^{td(y')} & 0 \\ 0 & 0_n \end{bmatrix}, \quad U_1(t) = \begin{bmatrix} 0 & 0 \\ 0 & e^{tD_1(y')} \end{bmatrix}.$$

Then, if we take a sufficiently small neighborhood V' of $y' = Y^{0\prime}$, there exist α, σ, $C > 0$, independent of y', such that in V',

$$|U_0(t)| \leqq e^{-(\alpha+\sigma)t} \quad (t \geqq 0), \tag{6.10}$$

$$|U_1(t)| \leqq C e^{\sigma t} \quad (t \leqq 0). \tag{6.11}$$

Theorems 4.1 and 4.2 of [2, Chap. 18] shows that there exist C^1-functions $r_i(w_0)$ near the origin in $\mathbb{R}$ satisfying $r_i(0) = r_i'(0) = 0$, and a solution w of (6.9) such that

$$b\, e^{td(y')} \leqq |w_0(t, y')| \leqq c\, e^{td(y')}, \tag{6.12}$$

$$w_i(t, y') = r_i(w_0(t, y')) \quad (1 \leqq i \leqq n), \tag{6.13}$$

for some constants b, $c > 0$. Furthermore we will show that w is C^1 with respect to y'. See the Appendices. Thus we conclude that $u = e^{-at}P(y')w$ is also C^1 with respect to (t, y').

LEMMA 13. *For any $\varepsilon > 0$, there exists a small neighborhood V' of $y' = Y^{0\prime}$ such that*

$$u_0(t, y') \sim e^{-(\alpha - d(y'))t}, \tag{6.14}$$

$$|u_i(t, y')| \leqq \varepsilon |u_0(t, y')| \quad (1 \leqq i \leqq n-1), \tag{6.15}$$

in V' as $t \to +\infty$.

Proof. We can easily obtain (6.14) from (6.12). We show (6.15). Recall that $u_i = e^{-at} \sum_{j=0}^{n} p_{ij}(y')w_j$. From (6.8) and (6.13), we have

$$u_0 \sim e^{-at} w_0,$$

$$|u_i| \leqq \varepsilon\, e^{-at} |w_0| \quad (1 \leqq i \leqq n-1).$$

This completes the proof. □

Thus we obtain a family of integral curves of $\tilde{X}$.

LEMMA 14. *There exists a constant $c > 0$ such that*

$$\left| \det \frac{\partial(T, Y')}{\partial(t, y')} \right| \geqq c |u_0(t, y')|. \tag{6.16}$$

Proof. A direct calculation implies

$$\frac{\partial(T, Y')}{\partial(t, y')} = \begin{bmatrix} \partial u_0/\partial t & * \\ \partial u_1/\partial t & \\ \vdots & I_{n-1} + [\partial u_i/\partial y_j]_{1 \leqq i,j \leqq n-1} \\ \partial u_{n-1}/\partial t & \end{bmatrix}. \tag{6.17}$$

Since $\partial u_0/\partial t \sim u_0$, $|\partial u_i/\partial t| \leqq \varepsilon |u_0|$ $(1 \leqq i \leqq n-1)$, we get the conclusion (6.16). This completes the proof. □

Now we conclude that the orbits of $\tilde{X}$ flow from a C^1-hypersurface in $\mathbb{R}^{n+1}_{(T,Y)}$ parametrized by (T, Y'), which we write $\tilde{S}=\{Y_n=\tilde{\psi}(T, Y')\}$.

LEMMA 15. *$\partial\tilde{\psi}/\partial Y_j$ $(1\leqq j\leqq n-1)$ are bounded.*

Proof. Let Ψ' be the C^1-mapping: $\mathbb{R}^n_{(t,y')}\to\mathbb{R}^n_{(T,Y')}$ defined by $\Psi'(t, y')=(T(t, y'), Y'(t, y'))$. Then $\tilde{\psi}(T, Y')=Y_n\circ\Psi'^{-1}(T, Y')$, and $\partial\tilde{\psi}/\partial Y_j=\partial Y_n/\partial t\cdot\partial t/\partial Y_j+\sum_{i=1}^{n-1}\partial Y_n/\partial y_i\cdot\partial y_i/\partial Y_j$. It easily follows from (6.17) that $\partial y_i/\partial Y_j$ are bounded and $\partial t/\partial Y_j=O(u_0^{-1})$. Since $\partial Y_n/\partial t=O(u_0)$, $\partial\tilde{\psi}/\partial Y_j$ are bounded. This completes the proof. □

Now the function $\tilde{\nu}(T, Y)=Y_n-\tilde{\psi}(T, Y')$ satisfies $\tilde{X}\tilde{\nu}=0$ and $\tilde{\nu}|_{\Sigma^{1,1}}=0$. Moreover, since $|u_i|\leqq\varepsilon|u_0|$ $(1\leqq i\leqq n-1)$, $|u_n|\leqq C|u_0|$ on $\tilde{\nu}=0$, $1+T\lambda\neq 0$ on $\tilde{\nu}=0$. Let $\nu(t, x)=\tilde{\nu}\circ(\tilde{\Phi}|_{\tilde{R}_2})^{-1}(t, x)$. Then ν satisfies (4.2)-(4.3). By the mapping $\tilde{\Phi}$, $\tilde{S}=\{Y_n=\tilde{\psi}(T, Y')\}$ is transformed into $S=\{x_n=\psi(t, x')\}$. Here ψ is a C^1-function near $(t^0, x^{0\prime})$. S is the desired shock surface. Lemma 15 implies that $\partial\psi/\partial x_i$ $(1\leqq i\leqq n-1)$ are bounded. Thus we obtain the following theorem.

THEOREM. *Suppose* (A.1) *and* (A.2) *are true. Then the function u constructed above is the entropy solution for* (1.1) *having C^1-shock surface near (t^0, x^0).*

7. An example. Consider the following equation in two space dimensions:

$$\frac{\partial u}{\partial t}+\sum_{i=1}^{2}\frac{\partial}{\partial x_i}(v_i f(u))=0.$$

This equation arises in oil reservoir problems; see [5] and [16]. We assume that v_i are positive constants and that $f(u)=u^2/2$. We treat the Cauchy problem with the following initial condition:

$$u(0, x)=\varphi(x)=(x_1^3+x_2^3)/3-(x_1+x_2).$$

Then we have

$$\Phi_t\colon x_i=y_i+tv_i\varphi(y), \qquad i=1, 2$$

$$\lambda(y)=\sum_{i=1}^{2}v_i(y_i^2-1),$$

$$\min\lambda(y)=\lambda(0)=-(v_1+v_2)<0.$$

$$\text{Hess }\lambda=\text{diag}\,(2v_1, 2v_2),$$

$$\Sigma^1=\{v_1y_1^2+v_2y_2^2=v_1+v_2-1/t\},$$

$$\Sigma^{1,1}=\{(t, y)\in\Sigma^1;\ v_1^2y_1+v_2^2y_2=0\}.$$

Especially, if we set $v_1=v_2=1$, then,

$$\Sigma^1=\{y_1^2+y_2^2=2-1/t\}, \qquad \Sigma^{1,1}=\{y_1^2+y_2^2=2-1/t,\ y_1+y_2=0\}$$

and we can easily construct the entropy solution explicitly, whose shock surface is expressed by $x_1+x_2=0$, $|x_1|\leqq 1-1/2t$.

Appendices.

A1. The stable manifold theory.

In the Appendices, we will show the smoothness of the solution w of (6.9) with respect to the parameter y'. First we will survey the stable manifold theory as preliminaries in this section. For the details, see [2].

LEMMA 16. *A function w is a bounded solution of* (6.9) *if and only if it is a bounded solution of the following*:

$$\begin{aligned} w(t) &= U_0(t)\xi + \int_0^t U_0(t-s)g(s, y', w(s))\,ds \\ &\quad - \int_t^\infty U_1(t-s)g(s, y', w(s))\,ds \\ &\quad (= T_\xi w) \end{aligned} \tag{A1.1}$$

for some $\xi \in \mathbb{R}^{n+1}$.

We prepare some notation:

$$X_k = \{w(t) \in C^0([0,\infty)); |w|_k = \sup_{t\geqq 0} e^{\alpha k t}|w(t)| < \infty\},$$

$$B_k(r) = \{u \in X_k; |u|_k \leqq r\} \quad (k = 0, 1),$$

$$\gamma(\rho) = \sup\{|\nabla_w g(t, y', w)|, |\nabla_w(\partial g/\partial y_j)(t, y', w)|;\ t \geqq 0,\ y' \in V',\ |w| \leqq \rho,\ 1 \leqq j \leqq n-1\}.$$

Note that $\gamma(\rho)$ is continuous near $\rho = 0$ and $\gamma(0) = 0$.

LEMMA 17. *Let* $C_1 = C_0 + 1$. *Then,*

$$|T_\xi w|_k \leqq |\xi| + \sigma^{-1}C_1\gamma(|w|_0)|w|_k,$$

$$|T_\xi w - T_\xi v|_k \leqq \sigma^{-1}\gamma(\max(|w|_0, |v|_0))|w - v|_k,$$

for $k = 0, 1$.

If we take ρ such that $\gamma(2\rho) < \sigma/(2C_1)$, then Lemma 17 implies, for $|\xi| \leqq \rho$, $k = 0, 1$,

$$T_\xi : B_k(2\rho) \to B_k(2\rho), \qquad |T_\xi w - T_\xi v|_k \leqq \tfrac{1}{2}|w - v|_k.$$

Thus we have the following lemma.

LEMMA 18. *Suppose* $|\xi| \leqq \rho$. *Then the equation* $T_\xi w = w$ *has a unique solution* $w = w(t, y', \xi)$ *in* $B_1(2\rho)$.

We can take $\xi = {}^t(\xi, 0, \cdots, 0)$, $\xi \in \mathbb{R}$. Set $w = {}^t(w_0, w')$ and $g = {}^t(g_0, g')$. Then (A1.1) implies:

$$w_0(0, y', \xi) = \xi,$$

$$\begin{aligned} w'(0, y', \xi) &= -\int_0^\infty U_1(-s)g'(s, y', w(s, y', \xi))\,ds \\ &\quad (= S(\xi)). \end{aligned}$$

LEMMA 19. $S(\xi)$ *is of* C^1 *near* $\xi = 0$ *and satisfies*:

$$|S(\xi)| \leqq \frac{2|\xi|}{\alpha + \sigma}\gamma(2|\xi|),$$

$$|S'(\xi)| \leqq \frac{2C_0}{\alpha + \sigma}\gamma(2|\xi|).$$

A2. The existence of $\partial w/\partial y_j$. We will show the smoothness of w with respect to y'. The method of proof is the same as in the proof of Theorem 4.2 in [2, Chap. 13].

Let us differentiate (A1.1) formally with respect to y_j. Then, $\eta = \partial w/\partial y_j$, if it exists, must satisfy:

$$\begin{aligned}\eta = t\frac{\partial d}{\partial y_j}U_0(t)\xi &+ \int_0^t (t-s)\frac{\partial d}{\partial y_j}U_0(t-s)g(s,y',w(s))\,ds\\ &+\int_0^t U_0(t-s)\frac{\partial g}{\partial y_j}(s,y',w(s))\,ds\\ &+\int_0^t U_0(t-s)\nabla_w g(s,y',w(s))\cdot\eta(s)\,ds\\ &-\int_t^\infty \frac{\partial D_1}{\partial y_j}(t-s)U_1(t-s)g(s,y',w(s))\,ds\\ &-\int_t^\infty U_1(t-s)\frac{\partial g}{\partial y_j}(s,y',w(s))\,ds\\ &-\int_t^\infty U_1(t-s)\nabla_w g(s,y',w(s))\cdot\eta(s)\,ds\\ &= I_1+I_2+\cdots+I_7.\end{aligned} \tag{A2.1}$$

We will show the existence of η satisfying (A2.1). Let us estimate I_j. We estimate I_2 for example. From (6.11) and the definition of γ, we have, for some $C>0$,

$$\begin{aligned}|I_2| &\leqq C\int_0^t (t-s)\,e^{-(\alpha+\sigma)(t-s)}|g(s,y',w(s))|\,ds\\ &\leqq C\,e^{-\alpha t}\gamma(|w|_0)|w|_1\int_0^t (t-s)\,e^{-\sigma(t-s)}\,ds\\ &\leqq \sigma^{-2}C\,e^{-\alpha t}\gamma(|w|_0)|w|_1.\end{aligned}$$

Then $|I_2|_1 \leqq \sigma^{-2}C\gamma(|w|_0)|w|_1$. The same argument shows there exists $C>0$ such that $|I_1+I_2+I_3+I_5+I_6|_1 \leqq C\gamma(|w|_0)|w|_1$ and $|I_4+I_7|_1 \leqq \sigma^{-1}C_1\gamma(|w|_0)|\eta|_1$. Set $I_4+I_7 = U\eta$. By taking $|\xi|<\rho$ sufficiently small, we have $|I_1+I_2+I_3+I_5+I_6|_1 \leqq 1$ and $|U\eta|_1 \leqq \frac{1}{2}|\eta|_1$. Hence there exists $(I-U)^{-1} = \sum_{j=0}^\infty U^j$ and $|(I-U)^{-1}|_1 \leqq 2$. Then $\eta = (I-U)^{-1}(I_1+I_2+I_3+I_5+I_6)$ is the unique solution of (A2.1) in X_1.

Next we set $w_h = w(t,y'+he_j,\xi)$, $U_{i,h}(t) = U_i(t,y'+he_j)$, $g_h(t,y',w) = g(t,y'+he_j,w)$, where $e_j \in \mathbb{R}^{n-1}$, $e_{ji} = \delta_{ji}$, and consider

$$\begin{aligned}w_h - w = \{U_{0,h}(t)-U_0(t)\}\xi &+ \int_0^t U_{0,h}(t-s)\{g_h(s,y',w_h)-g(s,y',w_h)\}\,ds\\ &+\int_0^t \{U_{0,h}(t-s)-U_0(t-s)\}g(s,y',w_h)\,ds\\ &+\int_0^t U_0(t-s)\{g(s,y',w_h)-g(s,y',w)\}\,ds\\ &-\int_t^\infty U_{1,h}(t-s)\{g_h(s,y',w_h)-g(s,y',w_h)\}\,ds\\ &-\int_t^\infty \{U_{1,h}(t-s)-U_1(t-s)\}g(s,y',w_h)\,ds\\ &-\int_t^\infty U_1(t-s)\{g(s,y',w_h)-g(s,y',w)\}\,ds\\ &= I_1+\cdots+I_7.\end{aligned}$$

A direct calculation shows that there exists $C>0$ such that

$$|I_1+I_2+I_3+I_5+I_6|_1 \leqq C|h|,$$

$$|I_4|_1 \leqq \sigma^{-1}\gamma(\max(|w_h|_0, |w|_0))|w_h-w|_1,$$

$$|I_7|_1 \leqq (\alpha+\sigma)^{-1}C_0\gamma(\max(|w_h|_0, |w|_0))|w_h-w|_1.$$

Hence, if $|w_h|_0, |w|_0 \leqq 2|\xi|$, there exists $C>0$ such that $|w_h-w|_1 \leqq C|h|$.

Now, consider $\varepsilon=(w_h-w)/h-\eta$. Then,

$$\begin{aligned}
\varepsilon = &\frac{1}{h}\{U_{0,h}(t)-U_0(t)\}\xi-\frac{\partial U_0}{\partial y_j}(t)\xi+\int_0^t U_{0,h}(t-s)\left\{\frac{1}{h}(g_h(w_h)-g(w_h))-\frac{\partial g}{\partial y_j}(w_h)\right\}ds\\
&+\int_0^t U_{0,h}(t-s)\left\{\frac{\partial g}{\partial y_j}(w_h)-\frac{\partial g}{\partial y_j}(w)\right\}ds\\
&+\int_0^t \{U_{0,h}(t-s)-U_0(t-s)\}\frac{\partial g}{\partial y_j}(w)\,ds\\
&+\int_0^t \left\{\frac{1}{h}(U_{0,h}(t-s)-U_0(t-s))-\frac{\partial U_0}{\partial y_j}(t-s)\right\}g(w_h)\,ds\\
&+\int_0^t \frac{\partial U_0}{\partial y_j}(t-s)\{g(w_h)-g(w)\}\,ds\\
&+\int_0^t U_0(t-s)\left\{\frac{1}{h}(g(w_h)-g(w))-\nabla_w g(w)\cdot\eta\right\}ds\\
&-\int_t^\infty U_{1,h}(t-s)\left\{\frac{1}{h}(g_h(w_h)-g(w_h))-\frac{\partial g}{\partial y_j}(w_h)\right\}ds\\
&-\int_t^\infty U_{1,h}(t-s)\left\{\frac{\partial g}{\partial y_j}(w_h)-\frac{\partial g}{\partial y_j}(w)\right\}ds\\
&-\int_t^\infty \{U_{1,h}(t-s)-U_1(t-s)\}\frac{\partial g}{\partial y_j}(w)\,ds\\
&-\int_t^\infty \left\{\frac{1}{h}(U_{1,h}(t-s)-U_1(t-s))-\frac{\partial U_1}{\partial y_j}(t-s)\right\}g(w_h)\,ds\\
&-\int_t^\infty \frac{\partial U_1}{\partial y_j}(t-s)\{g(w_h)-g(w)\}\,ds\\
&-\int_t^\infty U_1(t-s)\left\{\frac{1}{h}(g(w_h)-g(w))-\nabla_w g(w)\cdot\eta\right\}ds\\
=&\,I_1+\cdots+I_{13}.
\end{aligned}$$

We estimate I_2 for example. Since

$$I_2=\int_0^t U_{0,h}(t-s)\left\{\int_0^1\int_0^1 rh\frac{\partial^2 g}{\partial y_j^2}(s, y'+qrhe_j, w_h)\,dq\,dr\right\}ds,$$

$$\left|\frac{\partial^2 g}{\partial y_j^2}(s, y'+qrhe_j, w_h)\right| \leqq C|w_h|,$$

for some $C>0$, we have $|I_2|_1 \leqq \sigma^{-1}C|w_h|_1|h|$. Similar estimates follow for $I_1 \sim I_6$, $I_8 \sim I_{12}$. Now we estimate I_7. Since

$$I_7 = \int_0^t U_0(t-s)\left[\int_0^1 \nabla_w g((1-r)w+rw_h)\,dr\cdot\left\{\frac{1}{h}(w_h-w)-\eta\right\}\right.$$
$$\left.+\int_0^1 \{\nabla_w g((1-r)w+rw_h)-\nabla_w g(w)\}\,dr\cdot\eta\right]ds,$$

we have for some C, $C'>0$,

$$|I_7|_1 \leqq \sigma^{-1}\gamma(\max(|w_h|_0, |w|_0))|\varepsilon|_1 + C|w_h-w|_1|\eta|_1$$
$$\leqq d^{-1}\gamma(\max(|w_h|_0, |w|_0))|\varepsilon|_1 + C'|h|.$$

In the same way, we have

$$|I_{13}|_1 \leqq (\alpha+\sigma)^{-1}C_0\gamma(\max(|w_h|_0, |w|_0))|\varepsilon|_1 + C'|h|.$$

Thus we obtain

$$|\varepsilon|_1 \leqq \sigma^{-1}C_1\gamma(\max(|w_h|_0, |w|_0))|\varepsilon|_1 + C|h| \leqq \tfrac{1}{2}|\varepsilon|_1 + C|h|.$$

Then $|\varepsilon|_1 \leqq 2C|h|$ and $\lim_{h\to 0}|\varepsilon|_1 = 0$. Hence $\partial w/\partial y_j$ exists and equals η. Similar argument shows the continuity of $\partial w/\partial y_j$.

Acknowledgments. The author would like to express his hearty gratitude to Professor M. Tsuji for his kind advice and encouragement. He also thanks Professor T. Ishii for helpful discussions.

REFERENCES

[1] N. M. CHEN, *On types of singularities for solutions of nonlinear hyperbolic systems*, Bull. Inst. Math. Acad. Sinica, 10 (1982), pp. 405-416.
[2] E. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
[3] E. D. CONWAY, *The formation and decay of shocks for a conservation law in several dimensions*, Arch. Rational Mech. Anal., 64 (1977), pp. 47-57.
[4] T. DEBENEIX, *Certains systèmes hyperboliques quasi-linéaires*, preprint.
[5] J. GLIMM, D. MARCHESIN, AND O. MCBRYAN, *Unstable fingers in two phase flow*, Commun. Pure Appl. Math., 34 (1981), pp. 53-75.
[6] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Graduate Texts in Mathematics, 14, Springer-Verlag, New York, 1973.
[7] J. GUCKENHEIMER, Lecture Notes in Mathematics, *Solving a single conservation law*, 468, Springer-Verlag, New York, 1975, pp. 108-134.
[8] ———, *Shocks and rarefactions in two space dimensions*, Arch. Rat. Mech. Anal., 59 (1975), pp. 281-291.
[9] P. F. HSIEH AND Y. SIBUYA, *A global analysis of matrices of functions of several variables*, J. Math. Anal. Appl., 14 (1966), pp. 332-340.
[10] G. JENNINGS, *Piecewise smooth solutions of a single conservation law exist*, Adv. in Math., 33 (1979), pp. 192-205.
[11] S. N. KRUZKOV, *First order quasilinear equations in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217-243.
[12] B. MORIN, *Formes canoniques de singularités d'une application différentiable*, C. R. Acad. Sci. Paris, 260 (1965), pp. 5662-5665.
[13] D. SCHAEFFER, *A regularity theorem for conservation laws*, Adv. in Math., 11 (1973), pp. 358-386.
[14] M. TSUJI, *Formation of singularities for Hamilton-Jacobi equation* I, Proc. Japan Acad., 59 (1983), pp. 55-58.

[15] M. TSUJI, *Formation of singularities for Hamilton-Jacobi equation* II, J. Math. Kyoto Univ., 26 (1986), pp. 299-308.

[16] D. WAGNER, *The Riemann problem in two space dimensions for a single conservation law*, SIAM J. Math. Anal., 14 (1983), pp. 534-559.

[17] H. WHITNEY, *On singularities of mappings of Euclidean space* I, *Mappings of the plane into the plane*, Ann. of Math., 62 (1955), pp. 374-410.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

012

ON THE STRONGLY DAMPED WAVE EQUATION: $u_{tt}-\Delta u-\Delta u_t+f(u)=0$*

DANG DINH ANG† AND ALAIN PHAM NGOC DINH‡

Abstract. The following initial boundary value problem is considered

$$u_{tt}-\Delta u-\lambda\Delta u_t+f(u)=0,\quad (x,t)\in\Omega\times]0,T[,\quad \lambda>0,$$
$$u=0\quad \text{on } \partial\Omega\times[0,T),$$
$$u(x,0)=w_0(x),\qquad u_t(x,0)=w_1(x)$$

where Ω is a bounded domain in R^N with a sufficiently regular boundary $\partial\Omega$. In Part 1, a theorem on local existence and uniqueness is proved for w_0 in $H_0^1(\Omega)$ and w_1 in $L^2(\Omega)$, under a certain Lipschitzian condition on f.

In Part 2, the question of global existence and asymptotic behavior for $t\to\infty$ is studied, under more restrictive conditions, namely $1\leqq N\leqq 3$, $f\in C^1(\mathbb{R},\mathbb{R})$, $f(0)=0$, and $f'\geqq -c$ with $c>0$ "small" and $w_0\in H_0^1(\Omega)\cap H^2(\Omega)$, $w_1\in L^2(\Omega)$. It is proved that under these conditions, a unique solution $u(t)$ exists on $\mathbb{R}_+$ such that $\|u_t(t)\|$ and $\|\Delta u(t)\|$ decay exponentially to 0 as $t\to\infty$. ($\|\cdot\|$ denotes the $L^2(\Omega)$ norm.) The method followed in this paper is that of successive linearizations (Part 1) and Galerkin (Part 2).

Key words. linear recursive schemes, local existence, global existence, decay exponentially

AMS(MOS) subject classifications. 35, 35B, 35K, 35L, 41

Introduction. We will consider the following initial boundary value problem:

$$u_{tt}-\Delta u-\lambda\Delta u_t+f(u)=0,\quad (x,t)\in\Omega\times(0,T),\quad \lambda>0, \tag{0.1}$$

$$u=0\quad \text{on } \partial\Omega\times[0,T), \tag{0.2}$$

$$u(x,0)=w_0(x),\qquad u_t(x,0)=w_1(x), \tag{0.3}$$

where Ω is a bounded in $\mathbb{R}^N$ with a sufficiently regular boundary $\partial\Omega$. The problem was considered by Webb in [9] for $N=1,2,3$. For $\lambda=0$ and $N=1$, the problem was treated in [7] and [8], and in [2] for a function f depending on u and u_t or monotone in u_t. We consider the problem with $\lambda>0$ under various conditions on f and on the initial values. For $1\leqq N\leqq 2$, (0.1) governs the motion of a linear Kelvin solid (a bar if $N=1$ and a plate if $N=2$) subjected to nonlinear elastic constraints.

The paper consists of two parts. In Part 1 under a certain local Lipschitzian condition on f with w_0 in $H_0^1(\Omega)$ and w_1 in $L^2(\Omega)$, a local existence and uniqueness theorem is proved, using the method of successive linearizations. Some of the results on local existence are also contained in [1] and [4]. In [4] Sandefur factors (0.1) and then uses semigroups and successive approximations to get existence and uniqueness. The results in [1] are generalizations of work done in [4].

In Part 2, we strengthen the hypotheses and assume as in [9] that $1\leqq N\leqq 3$, $w_0\in H_0^1(\Omega)\cap H^2(\Omega)$, and $w_1\in L^2(\Omega)$ while f satisfies no condition other than $f(0)=0$, $f'\geqq -c$ for $c>0$ "small." It is then proved that under these conditions, a unique solution $u(t)$ exists for all $t\geqq 0$, with the property that $\|u_t(t)\|$ and $\|\Delta u(t)\|$ decay exponentially to 0 as $t\to\infty$ generalizing a result of Webb [9]. (Here and elsewhere $\|\cdot\|$

* Received by the editors September 17, 1986; accepted for publication (in revised form) February 21, 1988.

† Ho-Chi-Minh City University, Ho Chi Minh City, Vietnam.

‡ Département de Mathématiques et d'Informatique, Université d'Orléans, 45067 Orléans Cedex 2, France.

stands for the $L^2(\Omega)$ norm.) In Part 2 we limit ourselves to the case $1 \leqq N \leqq 3$ in order to use the imbedding theorem of Sobolev: $H^2(\Omega) \subset C(\bar{\Omega})$. Asymptotic results for strongly damped wave equations can also be found in [5]. In [3] we study an extension of equation (0.1), namely the strong solutions for an operator generalizing the Laplacian. The method used in Part 1 can serve as a starting point for computing algorithms [6]. Note that in this sequel we will consider (0.1) as an ordinary differential equation in a Banach space for $u(t)$, so that we will write $u'(t)$ for $u_t(\cdot, t)$. For simplicity of writing, we will take $\lambda = 1$.

1. Part 1. Let

$$L^2 = L^2(\Omega), \quad H_0^1 = H_0^1(\Omega), \quad H^2 = H^2(\Omega).$$

Here $H_0^1(\Omega)$ and $H^2(\Omega)$ denote the usual Sobolev spaces on Ω. Let $(\cdot)$ denote either the L^2 inner product or the paring of a continuous linear functional with an element of a function space. Let $\|\cdot\|_x$ be a norm on a Banach space X, and let X^* be its dual. We denote by $L^p(0, T; X)$, $1 \leqq p \leqq \infty$, the space of functions f on $(0, T)$ to X such that

$$\|f\|_{L^p(0,T;X)} = \left(\int_0^T \|f(t)\|_x^p \, dt \right)^{1/p} < +\infty \quad \text{for } 1 \leqq p < \infty,$$

$$\|f\|_{L^\infty(0,T;X)} = \operatorname*{ess\,sup}_{(0,T)} \|f(t)\|_x \quad \text{for } p = \infty.$$

We will make the following assumption

(A_1) $\qquad f: H_0^1 \to H^{-1}$ satisfies:

for each bounded subset B of $H_0^1(\Omega)$, there exists $k_B > 0$ such that

$$\|f(y) - f(z)\|_{H^{-1}} \leqq k_B \|\nabla y - \nabla z\| \quad \forall y, z \in B$$

where $\|\cdot\|_{H^{-1}}$ is the dual norm on H^{-1}, the dual of $H_0^1(\Omega)$.

Then we have the following theorem.

THEOREM 1. *Suppose f satisfies (A_1) and let $w_0 \in H_0^1$, $w_1 \in L^2$. Then there exists a $T > 0$ such that the initial and boundary value problem (0.1)-(0.3) admits a unique solution $u(t)$ in the following sense*:

(i) $u \in C(0, T; H_0^1)$,

(ii) $u' \in C(0, T; L^2) \cap L^2(0, T; H_0^1)$,

(iii) $d/dt(u'(t), v) + a(u'(t), v) + a(u(t), v) + (f(u(t)), v) = 0 \quad \forall v \in H_0^1$,

$u(0) = w_0, \quad u'(0) = w_1$

with $a(u, v) = (\nabla u, \nabla v)$.

Furthermore, $u(t)$ is the limit of the sequence $\{u_n(t)\}$ of solutions of the following initial boundary value problems (i.b.v. problem):

$$u_n'' - \Delta u_n' - \Delta u_n = -f(u_{n-1}), \quad n \geqq 1, \quad u_0 = 0, \tag{1.1}$$

$$u_n = 0 \quad \text{on } \partial\Omega, \tag{1.2}$$

$$u_n(0) = w_0, \quad u_n'(0) = w_1. \tag{1.3}$$

The sequence $\{u_n\}$ converges uniformly to u in $C(0, T; H_0^1)$ and the sequence $\{u_n'\}$ converges to u' in $L^2(0, T; H_0^1)$ and uniformly in $C(0, T; L^2)$.

The proof of the theorem relies on a number of propositions.

PROPOSITION 1. *Under assumption* (A_1), *there exists a* $T>0$ *such that for each* n *the initial boundary value problem* (1.1)–(1.3) *admits a unique solution* u_n *in* $C(0, T; H_0^1)$ *with* u_n' *in* $C(0, T; L^2) \cap L^2(0, T; H_0^1)$. *Furthermore the sequence* $\{u_n\}$ *lies in a bounded subset of* $C(0, T; H_0^1)$ *and the sequence* $\{u_n'\}$ *lies in a bounded subset of* $C(0, T; L^2) \cap L^2(0, T; H_0^1)$.

Proof. Consider the following equation in H_0^1 for $u_1(t)$:

$$u_1'' - \Delta u_1 - \Delta u_1' = -f(0) \tag{1.4}$$

with the initial conditions

$$u_1(0) = w_0, \qquad u_1'(0) = w_1. \tag{1.5}$$

It is easily proved, using a Galerkin approximation scheme, that for each $K>0$, a unique solution $u_1(t)$ exists on $[0, K]$. Suppose by induction that $u_{n-1}(t)$ is a solution of the i.b.v. problem

$$u_{n-1}'' - \Delta u_{n-1} - \Delta u_{n-1}' = -f(u_{n-2}), \tag{1.6}$$

$$u_{n-1}(0) = w_0, \qquad u_{n-1}'(0) = w_1 \tag{1.7}$$

satisfying

$$\|\nabla u_{n-1}(t)\|^2 + \|u_{n-1}'(t)\|^2 + \int_0^t \|\nabla u_{n-1}'(\tau)\|^2 \, d\tau \leqq M^2, \qquad 0 \leqq t \leqq T \tag{1.8}$$

where

$$M^2 > 3(\|\nabla w_0\|^2 + \|w_1\|^2), \tag{1.9}$$

$$T(k_M \cdot M + |f(0)|(\operatorname{mes}\Omega)^{1/2})^2 < 2M^2/3 \qquad (\operatorname{mes}\Omega = \text{Lebesgue measure of } \Omega), \tag{1.10}$$

k_M a constant >0 such that

$$\|f(y) - f(z)\|_{H^{-1}} \leqq k_M \|\nabla y - \nabla z\| \quad \forall y, z, \quad \text{with } \|\nabla y\| \leqq M, \quad \|\nabla z\| \leqq M. \tag{1.11}$$

We claim that the (unique) solution $u_n(t)$ of the i.b.v. problem

$$u_n'' - \Delta u_n - \Delta u_n' = -f(u_{n-1}), \tag{1.12}$$

$$u_n(0) = w_0, \qquad u_n'(0) = w_1 \tag{1.13}$$

satisfies Proposition 1.

The existence of a solution of (1.12), (1.13) can be proved using a Galerkin approximation scheme. Taking the inner product of (1.12) with u_n' and integrating with respect to the time variable from 0 to t give, after some rearrangements,

$$\begin{aligned} &\|u_n'(t)\|^2 + \|\nabla u_n(t)\|^2 + 2\int_0^t \|\nabla u_n'(\tau)\|^2 \, d\tau \\ &= \|w_1\|^2 + \|\nabla w_0\|^2 - 2\int_0^t (f(u_{n-1}(\tau)), u_n'(\tau)) \, d\tau. \end{aligned} \tag{1.14}$$

We have

$$\left| -2\int_0^t (f(u_{n-1}(\tau)), u_n'(\tau)) \, d\tau \right| \leqq \int_0^t \|f(u_{n-1}(\tau))\|_{H^{-1}}^2 \, d\tau + \int_0^t \|\nabla u_n'(\tau)\|^2 \, d\tau. \tag{1.15}$$

Therefore,

$$\left|-2\int_0^t (f(u_{n-1}(\tau)), u_n'(\tau))\,d\tau\right| \leqq T[k_M M + |f(0)|(\operatorname{mes}\Omega)^{1/2}]^2 + \int_0^t \|\nabla u_n'(\tau)\|^2\,d\tau \tag{1.16}$$

as can be seen from the induction hypothesis (1.8) and condition (A_1). From (1.16), (1.9), and (1.10) we can then deduce

$$\|u_n'(t)\|^2 + \|\nabla u_n(t)\|^2 + \int_0^t \|\nabla u_n'(\tau)\|^2\,d\tau \leqq M^2 \quad \forall 0 \leqq t \leqq T. \tag{1.17}$$

We conclude that the latter inequality holds for all n. The continuity of $u_n : [0, T) \to H_0^1$ results from the fact that, by (1.17),

$$\int_0^t \|\nabla u_n'(\tau)\|^2\,d\tau \leqq M^2, \qquad 0 \leqq t \leqq T. \tag{1.18}$$

The mapping: $t \to \|u_n'(t)\|$ which is defined by (1.14) is continuous. On the other hand, for each v in H_0^1 the mapping, $t \to (u_n'(t), v)$, is a continuous mapping as can be seen from

$$\begin{aligned}(u_n'(t), v) + \int_0^t (\nabla u_n(\tau), \nabla v)\,d\tau + (\nabla u_n(t), \nabla v)\\ = (w_1, v) - \int_0^t (f(u_{n-1}(\tau)), v)\,d\tau + (\nabla u_n(0), \nabla v)\end{aligned} \tag{1.19}$$

which is obtained from (1.12) by taking the inner product with v in H_0^1 and integrating with respect to the time variable from 0 to t. Hence u_n' is continuous on $[0, T)$ to L^2. □

PROPOSITION 2. *Let T satisfy* (1.10) *and furthermore let*

$$k_M^2 T < 1 \tag{1.20}$$

with M and k_M as in (1.9)-(1.11). *Then the sequence $\{u_n\}$ constructed in the proof of Proposition* 1 *is a Cauchy sequence in $C(0, T; H_0^1)$. Furthermore the sequence $\{u_n'\}$ is a Cauchy sequence in $C(0, T; L^2) \cap L^2(0, T; H_0^1)$.*

Proof. Let $v_n = u_n - u_{n-1}$. Then v_n satisfies

$$v_n'' - \Delta v_n - \Delta v_n' = -[f(u_{n-1}) - f(u_{n-2})], \tag{1.21}$$

$$v_n(0) = v_n'(0) = 0. \tag{1.22}$$

Taking the inner product with $v_n'(t)$, integrating, and rearranging gives

$$\begin{aligned}\|v_n'(t)\|^2 + \|\nabla v_n(t)\|^2 + \int_0^t \|\nabla v_n'(\tau)\|^2\,d\tau &\leqq k_M^2 \int_0^t \|\nabla v_{n-1}(\tau)\|^2\,d\tau\\ &\leqq k_M^2 T \sup_t \|\nabla v_{n-1}(t)\|^2\end{aligned} \tag{1.23}$$

where the sup is taken over $0 \leqq t \leqq T$.

It follows from (1.23) that

$$\sup_t \|\nabla v_n(t)\|^2 \leqq \sigma \sup_t \|\nabla v_{n-1}(t)\|^2 \quad \text{for } \sigma = k_M^2 T < 1.$$

Hence

$$\sup_t \|\nabla u_n(t) - \nabla u_m(t)\| \to 0 \quad \text{for } n, m \to \infty. \tag{1.24}$$

Furthermore we have for $n > m$

$$u_n'' - u_m'' - \Delta(u_n - u_m) - \Delta(u_n' - u_m') = -[f(u_{n-1}) - f(u_{m-1})]. \tag{1.25}$$

Taking the inner product with $u_n'(t)-u_m'(t)$, integrating with respect to t, and making use of the Lipschitzian property of f gives after some rearrangements

$$\begin{aligned}&\|u_n'(t)-u_m'(t)\|^2+\|\nabla u_n(t)-\nabla u_m(t)\|^2+\int_0^t \|\nabla u_n'(\tau)-\nabla u_m'(\tau)\|^2\,d\tau\\ &\leqq k_M^2\int_0^t \|\nabla u_{n-1}(\tau)-\nabla u_{m-1}(\tau)\|^2\,d\tau.\end{aligned}\tag{1.26}$$

It follows from (1.24) and (1.26) that

$$\sup_t \|u_n'(t)-u_m'(t)\|\to 0 \quad \text{for } n, m\to\infty \tag{1.27}$$

and

$$\int_0^T \|\nabla u_n'(\tau)-\nabla u_m'(\tau)\|^2\,d\tau\to 0 \quad \text{for } n, m\to\infty. \tag{1.28}$$

□

We turn to the proof of Theorem 1.

Proof of Theorem 1. It is immediate that there is at most one solution of (0.1)-(0.3). We will prove existence. Since H_0^1 is complete and since, by Proposition 2, u_n is a Cauchy sequence in $C(0, T; H_0^1)$, there is a u in $C(0, T; H_0^1)$ such that

$$\sup_t \|\nabla u_n(t)-\nabla u(t)\|\to 0 \quad \text{for } n\to\infty. \tag{1.29}$$

Since clearly

$$u_n'\to u' \quad \text{weak * in } L^\infty(0, T; L^2) \tag{1.30}$$

and

$$u_n'\to u' \quad \text{weakly in } L^2(0, T; H_0^1),$$

we also have by Proposition 2

$$\sup_t \|u_n'(t)-u'(t)\|\to 0 \quad \text{as } n\to\infty, \tag{1.31}$$

$$\int_0^T \|\nabla u_n'(\tau)-\nabla u'(\tau)\|^2\,d\tau\to 0 \quad \text{as } n\to\infty. \tag{1.32}$$

From (1.1), we have

$$\begin{aligned}&(u_n'(t), v)+\int_0^t (\nabla u_n'(\tau), \nabla v)\,d\tau+\int_0^t (\nabla u_n(\tau), \nabla v)\,d\tau\\ &=-\int_0^t (f(u_{n-1}(\tau)), v)\,d\tau+(w_1, v) \quad \forall v\in H_0^1.\end{aligned}\tag{1.33}$$

From (1.29), (1.31), (1.32) we obtain, using the Lipschitzian property of f, and passing to the limit as $n\to\infty$ in (1.33):

$$\begin{aligned}&(u'(t), v)+\int_0^t (\nabla u'(\tau), \nabla v)\,d\tau+\int_0^t (\nabla u(\tau), \nabla v)\,d\tau\\ &=-\int_0^t (f(u(\tau)), v)\,d\tau+(w_1, v) \quad \forall v\in H_0^1.\end{aligned}\tag{1.34}$$

It follows from (1.34) and (1.17) that $u(t)$ satisfies

$$d/dt(u'(t), v)+(\nabla u'(t), \nabla v)+(\nabla u(t), \nabla v)=-(f(u(t)), v)$$

(1.35) almost everywhere on $(0, T)$ and $\forall v \in H_0^1$

$$u(0)=w_0, \qquad u'(0)=w_1. \qquad \square$$

2. Part 2. We will consider the problem of global existence and asymptotic behavior for $t \to \infty$. To this end, we will limit ourselves, in what follows, to the case $1 \leqq N \leqq 3$, and furthermore, we will restrict the hypotheses on f and on the regularity of the initial data. Thus we will consider the following conditions on f:

(A_2) $$f \in C^1(\mathbb{R}, \mathbb{R}), \qquad f(0)=0,$$

(A_3) $$(f(u)+\varepsilon u)u \geqq 0 \quad \forall |u| \geqq a,$$

with $0<\varepsilon<1$ satisfying $\varepsilon\alpha^2<1$ where $\alpha>0$ such that

(2.1) $$\|u\| \leqq \alpha \|\nabla u\| \quad \text{and} \quad \|\nabla u\| \leqq \alpha \|\Delta u\| \quad \forall u \in H_0^1 \cap H^2,$$

(A_4) $$f' \geqq -c, \qquad c>0.$$

The problem (0.1)-(0.3) with $w_0 \in H_0^1 \cap H^2$, $w_1 \in L^2$ was studied by Webb [9] under conditions on f which are similar to (A_2)-(A_4). We will show that Webb's result [9] on asymptotic decay for $t \to \infty$ can be considerably strengthened.

PROPOSITION 3. *Let $w_0 \in H_0^1 \cap H^2$ and $w_1 \in L^2$ and let f satisfy (A_2)-(A_4). Then there is a unique solution $u(t)$ of the i.b.v. problem (0.1)-(0.3) defined on $[0, \infty)$. Moreover the quantity*

$$\|\Delta u(t)\|^2+\|u'(t)\|^2+\int_0^t \|\nabla u'(\tau)\|^2 \, d\tau$$

is bounded on compact subsets of $[0, \infty)$.

Proof. Let the approximated problem be, after rewriting $f(u)=g(u)-\varepsilon u$ with $\varepsilon>0$

(2.2) $$u_n''-\Delta u_n-\Delta u_n'+g(u_n)-\varepsilon u_n=0.$$

Here the finite-dimensional spaces considered in the Galerkin approximation are eigenspaces of the Laplacian.

Taking the inner product of (2.2) with $u_n'(t)$ and integrating from 0 to t, we obtain

(2.3) $$\|u_n'(t)\|^2+\|\nabla u_n(t)\|^2+2\int_0^t \|\nabla u_n'(\tau)\|^2 \, d\tau+2\int_\Omega \int_0^{u_n} g(u) \, du-\varepsilon\|u_n(t)\|^2$$
$$=-\varepsilon\|w_0\|^2+\|w_1\|^2+\|\nabla w_0\|^2+2\int_\Omega \int_0^{w_0} g(u) \, du.$$

Since $w_0 \in H_0^1 \cap H^2$ and $1 \leqq N \leqq 3$, for ε sufficiently small we obtain from (2.3)

(2.4) $$\|u_n'(t)\|^2+\tilde{\beta}\|\nabla u_n(t)\|^2+2\int_0^t \|\nabla u_n'(\tau)\|^2 \, d\tau+2\int_\Omega \int_0^{u_n} g(u) \, du \leqq M'$$

where $\tilde{\beta}=1-\varepsilon\alpha^2>0$ and M' a constant independent of t.

For any argument $u_n>0$

(2.5) $$\int_0^{u_n} g(u) \, du=\int_0^a g(u) \, du+\int_a^{u_n} g(u) \, du \geqq -\int_{-a}^a |g(u)| \, du$$

since the second integral is nonnegative due to (A_3), u_n being larger than a. A similar argument applies for negative arguments u_n.

Finally from (2.4) we can deduce the following inequality:

$$\|u_n'(t)\|^2+\tilde{\beta}\|\nabla u_n(t)\|^2+2\int_0^t\|\nabla u_n'(\tau)\|^2\,d\tau\leqq 2(\operatorname{mes}\Omega)\int_{-a}^{a}|g(u)|\,du+M'=M \tag{2.6}$$

with M constant independent of t.

Likewise, by taking the inner product of (2.2) (with $\varepsilon=c$) with $-\Delta u_n(t)$ and integrating from 0 to t give, after some rearrangements:

$$\begin{aligned}&\int_0^t\|\Delta u_n(\tau)\|^2\,d\tau+\frac{1}{2}\|\Delta u_n(t)\|^2+\int_0^t\left(\int_\Omega g'(u_n)|\nabla u_n|^2\,dx\right)d\tau\\ &=c\int_0^t\|\nabla u_n(\tau)\|^2\,d\tau+\int_\Omega u_n'\Delta u_n\,dx+\int_0^t\|\nabla u_n'(\tau)\|^2\,d\tau\\ &\quad+\frac{1}{2}\|\Delta w_0\|^2-\int_\Omega w_1\Delta w_0\,dx\quad\forall t\geqq 0.\end{aligned} \tag{2.7}$$

By (2.6), (2.7), and hypothesis (A_4) it follows that, for each $T>0$

$$\int_0^t\|\Delta u_n(\tau)\|^2+\frac{1}{2}\|\Delta u_n(t)\|^2\leqq M_T+\|u_n'(t)\|\cdot\|\Delta u_n(t)\|\quad\forall 0\leqq t\leqq T \tag{2.8}$$

where M_T is a constant depending on T.

Inequalities (2.8) and (2.6) involve

$$\|\Delta u_n(t)\|\leqq M_T\quad\forall 0\leqq t\leqq T, \tag{2.9}$$

M_T always indicating a bound depending on T.

Inequalities (2.6) and (2.9) show that from the sequence $\{u_n\}$ we can deduce a subsequence still denoted $\{u_n\}$ which converges weak * to an element $u\in L^\infty(0,T;H_0^1\cap H^2)$ such that $u'\in L^\infty(0,T;L^2)\cap L^2(0,T;H_0^1)$.

Consider the sequence $\{f(u_n)\}$. We have

$$(f(u_n(t)),v)\to(f(u(t)),v)\quad\forall v\in H_0^1\quad\text{in } L^\infty(0,T)\text{ weak *} \tag{2.10}$$

since $f\in C^1(\mathbb{R},\mathbb{R})$ and u_n converges to u almost everywhere on $Q=(0,T)\times\Omega$.

If we pass to the limit in the variational form associated with the approximated equation (2.2), we find that u satisfies equation (1.35).

Finally for each $T>0$ there exists a unique solution $u(t)$, $0\leqq t<T$ of the i.b.v. problem (0.1)-(0.3) with u in $L^\infty(0,T;H_0^1\cap H^2)$ and u' in $L^\infty(0,T;L^2)\cap L^2(0,T;H_0^1)$ and such that

$$\|\Delta u(t)\|^2+\|u'(t)\|^2+\int_0^T\|\nabla u'(\tau)\|^2\,d\tau\quad\text{is bounded on } [0,T). \qquad\square$$

Remark 1. As shown in Webb [9] the solution $u(t)$ is actually more regular with respect to t than has been asserted in Proposition 3 above.

Remark 2. Under hypotheses analogous to the ones in Proposition 3 above, Webb [9, Thm. 3.1] claims that there exists a global bound on $\|\Delta u(t)\|^2+\|u'(t)\|^2$ for all $t\geqq 0$. However his proof is not valid, since the inequality [9, (3.18)] is not proved. On the other hand, on the question of global bound we have the following proposition.

PROPOSITION 4. *If $f\in C^1(\mathbb{R},\mathbb{R})$, satisfies $\underline{\lim}_{|x|\to\infty}f(x)$, $x\geqq 0$, then under the sole condition $w_0\in H_0^1$ and $w_1\in L^2$, there exists a global bound $\|u'(t)\|$ and $\|\nabla u(t)\|$ for all $t\geqq 0$.*

Proof. Let $g(u)=f(u)+\varepsilon u$ ($\varepsilon>0$). By using the same arguments as in Proposition 3 it is possible to obtain, for ε sufficiently small, a constant M' independent of t such that

$$\|u'(t)\|^2+\tilde{\beta}\|\nabla u(t)\|^2+2\int_0^t\|\nabla u'(\tau)\|^2\,d\tau+2\int_\Omega\int_{w_0}^{u}g(u)\,du\leqq M'. \tag{2.11}$$

The condition $\underline{\lim}_{|x|\to\infty}f(x)x\geqq 0$ can be written as $\underline{\lim}_{|u|\geqq a}ug(u)\geqq 0$ for a sufficiently large, and we can always assume that $a\geqq|w_0(x)|$ almost everywhere $x\in\Omega$.

We can prove as in Proposition 3 that

$$\left|\int_\Omega\int_{w_0}^{u}g(u)\,du\right|\leqq(\operatorname{mes}\Omega)\int_{-a}^{a}|g(u)|\,du$$

which implies (2.6) for all $t\geqq 0$. □

We can now state the main result of Part 2.

THEOREM 2. *Let $w_0\in H_0^1\cap H^2$ and $w_1\in L^2$. Let* (A_2) *hold and let*

$$f'\geqq -c,\quad c \text{ satisfying the following conditions: } 0<c<\tfrac{1}{2}, \tag{A_4'}$$

$$c\alpha^2<1\quad(\alpha \text{ as in } (2.1)).$$

Then the solution $u(t)$, which exists for all $t\geqq 0$ as per Proposition 3, decays exponentially to 0 as $t\to\infty$ in the following sense: there exists an $M>0$ and $\gamma>0$ such that

$$\|u'(t)\|^2+\|\Delta u(t)\|^2\leqq M e^{-\gamma t}\quad\forall t\geqq 0.$$

Proof. Let c be as in Theorem 2. We write

$$f(u)=g(u)-cu;$$

then $g'(u)\geqq 0$, and hence f satisfies (A_3). Thus by Proposition 3 the solution $u(t)$ exists on $[0,\infty)$.

If we take the inner product of (0.1) with $u'(t)$ and integrate with respect to the time variable from 0 to t, we find using the property $u.g(u)\geqq 0$ (see (2.3)) that there exists an $M_1>0$ independent of t such that

$$\|u'(t)\|^2+(1-c\alpha^2)\|\nabla u(t)\|^2+2\int_0^t\|\nabla u(\tau)\|^2\,d\tau\leqq M_1^2\quad\text{for all } t\geqq 0. \tag{2.12}$$

Likewise, by taking the inner product of (0.1) with $-\Delta u(t)$ and integrating from 0 to t, we get (see (2.7)), taking into account (2.12), two constants C_1 and C_2 independent of t such that

$$(1-c\alpha^2)\int_0^t\|\Delta u(\tau)\|^2\,d\tau+\tfrac{1}{2}\|\Delta u(t)\|^2\leqq C_1\|\Delta u(t)\|+C_2\quad\forall t\geqq 0. \tag{2.13}$$

Clearly (2.13) implies

$$\|\Delta u(t)\|\leqq M_2\quad\forall t\geqq 0. \tag{2.14}$$

M_2 is a constant independent of t.

Taking the inner product of (0.1) first with $u'(t)\,e^{\gamma t}$, then with $-\beta\Delta u\,e^{\gamma t}$ and integrating with respect to the time variable from 0 to t, we find, after rearranging and

summing up, and taking (2.14) into account:

$$
\begin{aligned}
&(1-c-2\beta)\, e^{\gamma t}\|u'(t)\|^2/2+(1-\alpha^2 c)\, e^{\gamma t}\|\nabla u(t)\|^2/2+\beta\, e^{\gamma t}\|\Delta u(t)\|^2/4\\
&\quad+[1-\beta-\gamma(1+\beta)\alpha^2/2]\int_0^t e^{\gamma t}\|\nabla u'(t)\|^2\, d\tau\\
&\quad+[-\gamma/2-c\beta-\gamma k'\alpha^2+\beta(1-\gamma)/\alpha^2]\int_0^t e^{\gamma\tau}\|\nabla u(\tau)\|^2\, d\tau\\
&\quad+e^{\gamma t}\int_\Omega G(u(x,t))\, dx+\beta\int_0^t g'(u)\|\nabla u(\tau)\|^2\, e^{\gamma\tau}\, d\tau\\
&\quad+(c\gamma/2)\int_0^t e^{\gamma\tau}\|u(\tau)\|^2\, d\tau\leqq C(w_0, w_1, \beta, \gamma)
\end{aligned}
\tag{2.15}
$$

where

$$
G(u)=\int_0^u g(z)\, dz, \tag{2.16}
$$

$$
k'=(\text{mes }(\Omega)/2)\operatorname*{Sup}_t |g'(\|u(t)\|_\infty)|. \tag{2.17}
$$

If in (2.15) we take $\beta=(1-2c)/2$ and

$$
\gamma=(1/\alpha^2)\min[(1+2c)2/(3-2c), (1-\alpha^2 c)(1-2c)/(1+k'\alpha^2+(1-2c)/\alpha^2)]
$$

then the coefficient of each term of the left-hand side of the inequality is positive and hence there exists an $M>0$ such that

$$
\|u'(t)\|^2+\|\Delta u(t)\|^2\leqq M\, e^{-\gamma t}\quad \forall t>0. \qquad \square
$$

Webb proved a very interesting result [9, Thm. 4.1] on the asymptotic behavior of the solution under the hypotheses

$$
f\in C^1(\mathbb{R},\mathbb{R}),\quad f(0)=0,\quad \underline{\lim}_{|x|\to\infty} f(x).\, x\geqq 0,\quad f'(x)\geqq -c\quad \text{where } c>0.
$$

From his theorem he deduced, under the foregoing hypotheses with an additional condition on the smallness of $c>0$, that $\|u'(t)\|$ and $\|\Delta u(t)\|$ tend to 0 as $t\to\infty$ [9, Cor. 4.1]. Our theorem is therefore a considerably stronger result than Webb's result of the asymptotic decay of the solution for $t\to\infty$, since on one hand the hypothesis $\underline{\lim}_{|x|\to\infty} f(x).\, x\geqq 0$ is not used and on the other hand our asymptotic decay is exponential.

Remark 3. As remarked by one of the referees, the restriction on c $(0<c<\frac{1}{2})$ can be removed by a scaling argument. In fact write (0.1) in terms of $\xi=\mu x$, $\tau=\mu t$. Then we have

$$
u_{\tau\tau}-\lambda\mu\Delta_\xi u_\tau-\Delta_\xi u+f(u)/\mu^2=0.
$$

Let $g(u)=f(u)/\mu^2$, with $f(u)$ always satisfying (A_4); then the previous conditions on c become

$$
0<c<\text{Min }(\mu^2/2, 1/\alpha^2)
$$

with $\mu>1/2\lambda$ (the coefficient of $e^{\gamma t}\|\Delta u(t)\|^2$ in (2.15) becomes $(\beta/4)(2\lambda\mu-1)$).

Acknowledgments. The authors wish to thank the referees for their constructive criticism and most pertinent remarks, leading to improvements in the original manuscript.

REFERENCES

[1] P. AVILES AND J. SANDEFUR, *Nonlinear second order equations with applications to partial differential equations*, Differential Equations, 58 (1985), pp. 404-427.

[2] D. D. ANG AND A. PHAM NGOC DINH, *Mixed problem for semi-linear wave equation with a non-homogeneous condition*, Nonlinear Analysis T.M.A., (1988).

[3] ———, *Strong solutions of a quasilinear equation with nonlinear damping*, SIAM J. Math. Anal., 19 (1988), pp. 337-347.

[4] J. SANDEFUR, *Existence and uniqueness of solutions of second order nonlinear differential equations*, SIAM J. Math. Anal., 14 (1983), pp. 477-487.

[5] P. MARCATI, *Stability for second order abstract evolution equations*, Nonlinear Analysis T.M.A., 8 (1984), pp. 237-252.

[6] E. L. ORTIZ AND A. PHAM NGOC DINH, *Linear recursive schemes associated with some nonlinear partial differential equations in one dimension and the Tau method*, SIAM J. Math. Anal., 18 (1987), pp. 452-464.

[7] A. PHAM NGOC DINH, *Sur un problème hyperbolique faiblement nonlinéaire en dimension* 1, Demonstratio Mathematica, 16 (1983), pp. 269-289.

[8] A. PHAM NGOC DINH AND NGUYÊN THANH LONG, *Linear approximation and asymptotic expansion associated to the nonlinear wave equation in one dimension*, Demonstratio Mathematica, 19 (1986), pp. 45-63.

[9] G. F. WEBB, *Existence and asymptotic behavior for a strongly damped nonlinear wave equation*, Canad. J. Math., 32 (1980), pp. 631-643.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

013

ON THE NODAL SETS OF THE EIGENFUNCTIONS OF THE STRING EQUATION*

CHAO-LIANG SHEN†

Abstract. In this paper the nodal sets of the eigenfunctions of the string equation are investigated. For n sufficiently large it is found that the shortest nodal domain of the nth eigenfunction must be one of the neighboring nodal domains of the maximum points, and the longest nodal domain of the nth eigenfunction must be one of the neighboring nodal domains of the minimum points of the density function of the string equation. A limit formula for the ratio of the longest length and the shortest length of the nodal domains of the nth eigenfunction is also proved, and some average formulae for the nodal domains are derived.

Key words. string equation, eigenvalues, eigenfunctions, nodal sets, nodal domains

AMS(MOS) subject classifications. 34B25, 34C10

Introduction. The subject of this paper is the investigation of the nodal sets of the eigenfunctions of the string equation

$$\begin{aligned} &y''(x)+\lambda\rho(x)y(x)=0, \qquad 0<x<\beta, \\ &y(0)=y(\beta)=0 \end{aligned} \tag{0.1}$$

where the *density function* $\rho(x)$ is a strictly positive C^2-function on the interval $[0, \beta]$. In § 1 we study the relation between the location of the nodal domains of extreme lengths of the nth eigenfunction $\varphi_n(x)$ of (0.1) and the extremal points of $\rho(x)$. We also prove a limit formula (Theorem 1.3) for the ratio of the lengths of the longest and the shortest nodal domains of φ_n. In § 2 we prove some average formulae for the lengths of nodal domains, and the nodal points of the eigenfunctions of (0.1).

1. Length of the nodal domains and the extreme values of the density function. Let $\varphi_n(x)$ be the nth eigenfunction of the string equation (0.1). We will denote $x_1^{(n)}<x_2^{(n)}<\cdots<x_{n-1}^{(n)}$ the nodal points of φ_n in the open interval $(0, \beta)$, and we denote $x_0^{(n)}=0$, $x_n^{(n)}=\beta$, $I_{n,j}=[x_{j-1}^{(n)}, x_j^{(n)}]$. Then, by the variational formula for λ_n, and the monotonicity principle for the comparison of the eigenvalues, we have

$$\begin{aligned} &\lambda_n=\int_{x_{j-1}^{(n)}}^{x_j^{(n)}}[\varphi_n'(x)]^2\,dx \Big/ \int_{x_{j-1}^{(n)}}^{x_j^{(n)}}\rho(x)[\varphi_n(x)]^2\,dx, \\ &\frac{\pi^2}{\max(\rho, I_{n,j})|I_{n,j}|^2}\leqq\lambda_n\leqq\frac{\pi^2}{\min(\rho, I_{n,j})|I_{n,j}|^2}, \end{aligned} \tag{1.1}$$

where $\max(\rho, I_{n,j})$ (respectively, $\min(\rho, I_{n,j})$) denotes the maximum (respectively, the minimum) of ρ in $I_{n,j}$, and $|I_{n,j}|$ denotes the length of the interval $I_{n,j}$.

LEMMA 1.1. *For $\delta>0$, there exists $n_0(\delta)$ such that*

$$|I_{n,k}|<\delta, \qquad k=1, 2, \cdots, n$$

for all $n\geqq n_0$.

* Received by the editors May 6, 1987; accepted for publication (in revised form) January 22, 1988. This research was supported in part by National Science Council grant 76-0208-M007-40 of the National Science Council of the Republic of China.

† Institute of Mathematics, National Tsing Hua University, Hsinchu, Taiwan 30043, Republic of China.

Proof. Suppose there exists a $\delta > 0$ such that there is a sequence n_j, $\lim_{j\to\infty} n_j = \infty$, and for each n_j there is a nodal domain I_{n_j,k_j} such that $|I_{n_j,k_j}| \geqq \delta$. Then, by (1.1)

$$\lambda_{n_j} \leqq \frac{1}{\min(\rho, [0,\beta])} \cdot \frac{\pi^2}{\delta^2},$$

which contradicts the fact that $\lim_{n\to\infty} \lambda_n = \infty$. □

In the rest of this section we will assume, for simplicity, that the density function $\rho(x)$ satisfies the following *condition* (A): *ρ has only finitely many critical points in* $[0,\beta]$, where the *critical points* of ρ are those points at which the derivative of ρ vanishes. We will call the values of ρ at its critical points the *critical values* of ρ.

We define the *neighboring nodal domains* of the extreme points of $\rho(x)$ as follows. Let x_* be one of the maximum (respectively, minimum) points of $\rho(x)$. If $x_* \in$ Interior $(I_{n,j})$, we call $I_{n,j-1}$, $I_{n,j}$, $I_{n,j+1}$ the neighboring nodal domains of x_*. If $x_* = x_j^{(n)}$, we call $I_{n,j}$, $I_{n,j+1}$ the neighboring nodal domains of x_*. Note that if $x_* = \beta$, the only neighboring nodal domain is $I_{n,n}$. If $x_* = 0$, $I_{n,1}$ is the neighboring nodal domain. The following theorem tells us the nodal domains of φ_n of extreme lengths have to be the neighboring nodal domains of the extreme points of $\rho(x)$.

THEOREM 1.2. *Suppose $\rho(x)$ satisfies condition* (A). *Then there exists n_0 such that for $n > n_0$ a shortest* (*respectively, a longest*) *nodal domain of the nth eigenfunction of* (0.1) *must be one of the neighboring nodal domains of the maximum* (*respectively, minimum*) *points of the density function $\rho(x)$.*

Proof. Let $\rho_1 > \rho_2 > \cdots > \rho_l$ (>0) be the sequence of the maximum value, the critical values, and the minimum value of ρ. Denote $\varepsilon_1 = \frac{1}{3}\min\{\rho_1-\rho_2, \rho_2-\rho_3, \cdots, \rho_{l-1}-\rho_l\}$, and define ε as follows:

$$\varepsilon = \begin{cases} \varepsilon_1 & \text{if } \rho_1 = \rho(\beta) = \rho(0), \\ \min\left\{\varepsilon_1, \dfrac{\rho_1-\rho(\beta)}{3}\right\} & \text{if } \rho_1 \neq \rho(\beta), \quad \rho(0) = \rho_1, \\ \min\left\{\varepsilon_1, \dfrac{\rho_1-\rho(0)}{3}\right\} & \text{if } \rho_1 = \rho(\beta), \quad \rho(0) \neq \rho_1, \\ \min\left\{\varepsilon_1, \dfrac{\rho_1-\rho(0)}{3}, \dfrac{\rho_1-\rho(\beta)}{3}\right\} & \text{if } \rho_1 \neq \rho(\beta), \quad \rho_1 \neq \rho(0). \end{cases}$$

Then, for this ε, there exists $\delta > 0$ such that $|\rho(x)-\rho(x')| < \varepsilon$ if $|x-x'| < \delta$. By Lemma 1.1, for this δ, there exists an n_0, such that for $n > n_0$, $|I_{n,k}| < \delta/2$, $k = 1, 2, \cdots, n$.

Suppose there is an $n > n_0$ such that one of the shortest nodal domains, say $I_{n,j}$, of φ_n is not a neighboring nodal domain of the maximum points of $\rho(x)$. We consider the three possible behaviors of $\rho(x)$ in $I_{n,j}$ separately: (1) $\rho(x)$ is strictly increasing in $I_{n,j}$; (2) $\rho(x)$ is strictly decreasing in $I_{n,j}$; (3) one of the local extreme points of ρ lies in the interior of $I_{n,j}$. Note that by the choice of ε, δ, if x is in one of the neighboring nodal domains of the maximum points of $\rho(x)$, then

$$\rho(x) > \rho_1 - \varepsilon \geqq \rho_1 - \frac{\rho_1-\rho_2}{3} = \frac{2\rho_1+\rho_2}{3}, \qquad \rho(x) > \rho_1 - \frac{\rho_1-\rho(\beta)}{3} = \frac{2\rho_1+\rho(\beta)}{3}$$

if $\rho_1 > \rho(\beta)$.

If case (3) happens, since $I_{n,j}$ is not a neighboring nodal domain of the maximum points of ρ, the maximum of the local extremal values of ρ in $I_{n,j}$ is less than or equal to ρ_2. Thus, by the choice of ε, δ, and the fact $|I_{n,j}| < \delta/2$, $\rho(x) < \rho_2 + \varepsilon \leqq \rho_2 + (\rho_1-\rho_2)/3 = (2\rho_2+\rho_1)/3$ for all x in $I_{n,j}$. Since $2\rho_2+\rho_1 < 2\rho_1+\rho_2$, for x in $I_{n,j}$, ξ

in any of the neighboring nodal domains of the maximum points of ρ we have $\rho(x)<\rho(\xi)$.

If case (1) happens, let $\xi_j=\sup\{x\in[0,\beta]\colon \rho \text{ is increasing in } [x_{j-1}^{(n)}, x]\}$. Then ξ_j is either a critical point of ρ in $(0,\beta)$, or is β itself. Suppose ξ_j is a critical point of ρ in $(0,\beta)$. If ξ_j is not a maximum point of ρ, then for all x in $I_{n,j}$ $\rho(x)\leqq\rho_2<(2\rho_1+\rho_2)/3<\rho(\xi)$ for all ξ in any of the neighboring nodal domains of the maximum points of ρ. If ξ_j is a maximum point of ρ, since $I_{n,j}$ is not a neighboring nodal domain of ξ_j, there is a neighboring nodal domain $I_{n,k}$ of ξ_j such that $\rho(x)<\rho(\xi)$ for all x in $I_{n,j}$, ξ in $I_{n,k}$. That is, if ρ is increasing in $I_{n,j}$, then there exists a neighboring nodal domain $I_{n,k}$ of the maximum points of ρ such that $\rho(x)<\rho(\xi)$ for all x in $I_{n,j}$, ξ in $I_{n,k}$. The same conclusion holds for the case $\xi_j=\beta$ and for case (2).

By the previous argument, we find that if $I_{n,j}$ is one of the shortest nodal domains of the nth eigenfunction φ_n, $n>n_0$, which is not a neighboring nodal domain of the maximum points of ρ, then we can find a neighboring nodal domain $I_{n,k}$ of the maximum points of ρ such that $\rho(x)<\rho(\xi)$ for all x in $I_{n,j}$, ξ in $I_{n,k}$. Since $|I_{n,j}|\leqq|I_{n,k}|$, for x in $I_{n,j}$, $x+x_{k-1}^{(n)}-x_{j-1}^{(n)}$ is in $I_{n,k}$. For x in $I_{n,j}$ let $\tilde{\rho}(x)=\rho(x+x_{k-1}^{(n)}-x_{j-1}^{(n)})$, $u(x)=\varphi_n(x+x_{k-1}^{(n)}-x_{j-1}^{(n)})$. Then

$$u''(x)+\lambda_n\tilde{\rho}(x)u(x)=0 \quad \text{in } I_{n,j}.$$

Note that $\tilde{\rho}(x)>\rho(x)$ for x in $I_{n,j}$ because $\rho(\xi)>\rho(x)$ for ξ in $I_{n,k}$, x in $I_{n,j}$. Since

$$\varphi_n''(x)+\lambda_n\rho(x)\varphi_n(x)=0 \quad \text{in } I_{n,j},$$
$$\varphi_n(x_{j-1}^{(n)})=\varphi_n(x_j^{(n)})=0,$$
$$\rho(x)<\tilde{\rho}(x) \quad \text{in } I_{n,j},$$

Sturm's comparison theorem (see [2, Thm. 3.1] or [1, § 10.3]) tells us that $u(x)$ has a zero ξ in the interior of $I_{n,j}$; i.e., the eigenfunction φ_n has a nodal point $\xi+x_{k-1}^{(n)}-x_{j-1}^{(n)}$,

$$x_{k-1}^{(n)}<\xi+x_{k-1}^{(n)}-x_{j-1}^{(n)}<x_k^{(n)},$$

which is absurd. Thus the shortest nodal domains of the nth eigenfunction φ_n, $n>n_0$, are among neighboring nodal domains of the maximum points of $\rho(x)$. A similar argument implies the longest nodal domains of φ_n are among neighboring nodal domains of the minimum points of $\rho(x)$. □

Theorem 1.2 has an interesting application.

THEOREM 1.3. *Let L_n, ℓ_n denote, respectively, the lengths of the longest and shortest nodal domains of the nth eigenfunction φ_n of* (0.1). *Then*

$$\lim_{n\to\infty}\left(\frac{L_n}{\ell_n}\right)^2=\frac{\max(\rho,[0,\beta])}{\min(\rho,[0,\beta])}. \tag{1.2}$$

Proof. By Theorem 1.2, if we let $I_{n,j_1(n)},\cdots,I_{n,j_{k_n}(n)}$ be the shortest (respectively, longest) nodal domains of φ_n, then $\max(\rho,[0,\beta])$ (respectively, $\min(\rho,[0,\beta])$) is the limit point of the sequence

$$\rho(x_{j_1(n_0+1)-1}^{(n_0+1)}),\ \rho(x_{j_1(n_0+1)}^{(n_0+1)}),\cdots,\rho(x_{j_{k_{n_0+1}}(n_0+1)-1}^{(n_0+1)}),$$
$$\rho(x_{j_{k_{n_0+1}}(n_0+1)}^{(n_0+1)}),\cdots,\rho(x_{j_1(n)-1}^{(n)}),\ \rho(x_{j_1(n)}^{(n)}),\cdots,$$
$$\rho(x_{j_{k_n}(n)-1}^{(n)}),\ \rho(x_{j_{k_n}(n)}^{(n)}),\cdots.$$

Let $I_{n,\ell(n)}$ be a nodal domain of φ_n of shortest length and let $I_{n,L(n)}$ be a nodal domain of φ_n of longest length. Then $\lim_{n\to\infty}\max(\rho,I_{n,\ell(n)})=\lim_{n\to\infty}\min(\rho,I_{n,\ell(n)})=\max(\rho,[0,\beta])$, $\lim_{n\to\infty}\max(\rho,I_{n,L(n)})=\lim_{n\to\infty}\min(\rho,I_{n,L(n)})=\min(\rho,[0,\beta])$, and

these limits are independent of the choice of $I_{n,\ell(n)}$, $I_{n,L(n)}$. Since, by (1.1), we have

$$\frac{\min(\rho, I_{n,\ell(n)})}{\max(\rho, I_{n,L(n)})} \leqq \frac{L_n^2}{\ell_n^2} \leqq \frac{\max(\rho, I_{n,\ell(n)})}{\min(\rho, I_{n,L(n)})}.$$

These inequalities and previous limit formulae imply (1.2). □

2. Some average formulae. Let $x_1^{(n)} < x_2^{(n)} < \cdots < x_{n-1}^{(n)}$ be the nodal points of the nth eigenfunction φ_n of (0.1) in the open interval $(0, \beta)$. Denote $x_0^{(n)} = 0$, $x_n^{(n)} = \beta$.

THEOREM 2.1.

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{j=1}^{n} \frac{1}{x_j^{(n)} - x_{j-1}^{(n)}} = \left\{ \int_0^\beta \rho(x)\, dx \right\} \left\{ \int_0^\beta \sqrt{\rho(x)}\, dx \right\}^{-2}. \tag{2.1}$$

Proof. By (1.1) we have

$$\begin{aligned} \lambda_n \sum_{j=1}^{n} \min(\rho, I_{n,j}) \cdot (x_j^{(n)} - x_{j-1}^{(n)}) &\leqq \sum_{j=1}^{n} \frac{\pi^2}{x_j^{(n)} - x_{j-1}^{(n)}} \\ &\leqq \lambda_n \sum_{j=1}^{n} \max(\rho, I_{n,j}) \cdot (x_j^{(n)} - x_{j-1}^{(n)}). \end{aligned} \tag{2.2}$$

Since $\lim_{n\to\infty} (\lambda_n / n^2) = \pi^2 / (\int_0^\beta \sqrt{\rho(x)}\, dx)^2$, (2.2) implies

$$\frac{\pi^2 \int_0^\beta \rho(x)\, dx}{(\int_0^\beta \sqrt{\rho(x)}\, dx)^2} = \pi^2 \lim_{n\to\infty} \frac{1}{n^2} \sum_{j=1}^{n} \frac{1}{x_j^{(n)} - x_{j-1}^{(n)}},$$

which is (2.1). □

Remark. We know that $(\int_0^\beta \sqrt{\rho(x)}\, dx)^2 \leqq \beta \int_0^\beta \rho(x)\, dx$, and the equality holds if and only if ρ is a nonnegative constant. If the limit of (2.1) is $1/\beta$, then $\rho(x)$ must be a constant. The following inverse spectral result is a consequence of Theorem 2.1: "If there are infinitely many eigenvalues λ_{n_j} of (0.1) such that the nodal domains of the corresponding eigenfunctions are of length β/n_j, then $\rho(x)$ is a constant."

It will be interesting to find the limit

$$\lim_{n\to\infty} \frac{1}{n^3} \sum_{j=1}^{n} \frac{1}{(x_j^{(n)} - x_{j-1}^{(n)})^2}. \tag{2.3}$$

If we follow the idea of the proof of Theorem 2.1, we find that in order to evaluate the limit (2.3), we have to be able to calculate the limit

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} \rho(\bar{x}_j^{(n)})}{n}, \tag{2.4}$$

where $\bar{x}_j^{(n)} \in [x_{j-1}^{(n)}, x_j^{(n)}]$ such that $\rho(\bar{x}_j^{(n)})$ is equal to the maximum (respectively, minimum) of $\rho(x)$ in the interval $[x_{j-1}^{(n)}, x_j^{(n)}]$. Since $\rho(x)$ is uniformly continuous on $[0, \beta]$, to evaluate the limit (2.4), it suffices to evaluate the following limit:

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} \rho(x_j^{(n)})}{n}. \tag{2.5}$$

In the course of studying the convergence of (2.3), we found that the limit (2.5) does exist, and can be evaluated explicitly as follows.

THEOREM 2.2.

$$\lim_{n\to\infty} \frac{\sum_{j=1}^{n} \rho(x_j^{(n)})}{n} = \frac{\int_0^\beta [\rho(x)]^{3/2}\, dx}{\int_0^\beta \sqrt{\rho(x)}\, dx}. \tag{2.6}$$

Theorem 2.2 is a corollary of the following interesting theorem.

THEOREM 2.3. *Let* $x_1^{(n)} < \cdots < x_n^{(n)}$ *be the zeros of the nth eigenfunction* $\varphi_n(x)$ *of* (0.1) *in* $(0, \beta]$. *Then for* $f \in C^1[0, \beta]$, *we have*

$$\lim_{n\to\infty} \frac{f(x_1^{(n)}) + \cdots + f(x_n^{(n)})}{n} = \int_0^\beta f(x)\sqrt{\rho(x)}\, dx \Big/ \int_0^\beta \sqrt{\rho(x)}\, dx. \tag{2.7}$$

Proof. Denote by $N_n(t)$ the number of points among $x_1^{(n)}, \cdots, x_n^{(n)}$, which lie in the interval $(0, t]$. Then by Problem 147 of [3, Part II] we have

$$f(x_1^{(n)}) + \cdots + f(x_n^{(n)}) = nf(\beta) - \int_0^\beta N_n(t) f'(t)\, dt. \tag{2.8}$$

We claim that

$$\lim_{n\to\infty} \frac{N_n(x)}{n} = \frac{\int_0^x \sqrt{\rho(t)}\, dt}{\int_0^\beta \sqrt{\rho(t)}\, dt}. \tag{2.9}$$

Then (2.7) follows immediately from (2.8) and (2.9).

Now we prove (2.9). We may assume φ_n is normalized so that $\int_0^\beta \varphi_n^2(x)\, dx = 1$. It is known (see [1, § 11.4], [2, p. 13]) that $\varphi_n(x)$ has the following asymptotic formula:

$$\varphi_n(x) = \left(\frac{2}{\pi}\right)^{1/2} \sin\left(n\pi \frac{\int_0^x \sqrt{\rho(t)}\, dt}{\int_0^\beta \sqrt{\rho(t)}\, dt}\right) + O\left(\frac{1}{n}\right). \tag{2.10}$$

We denote $\int_0^x \sqrt{\rho(t)}\, dt / \int_0^\beta \sqrt{\rho(t)}\, dt$ by $g(x)$. Equation (2.10) means that there exist constants K and n_0, such that

$$\left|\varphi_n(x) - \left(\frac{2}{\pi}\right)^{1/2} \sin(n\pi g(x))\right| \leqq \frac{K}{n},$$

for $n \geqq n_0$.

Choose $n_1 \geqq n_0$ such that $(K/n) < \frac{1}{3}(2/\pi)^{1/2}$ for $n \geqq n_1$. For $n \geqq n_1$, let $y_j^{(n)}$ be the point in $[0, \beta]$ such that

$$ng(y_j^{(n)}) = j - \tfrac{1}{2}, \qquad j = 1, 2, \cdots, n.$$

Note that for a fixed n, $y_1^{(n)}, \cdots, y_n^{(n)}$ are distinct because g is strictly increasing. For $n \geqq n_1$ we have

$$\left|\varphi_n(y_j^{(n)}) - \left(\frac{2}{\pi}\right)^{1/2} \sin(n\pi g(y_j^{(n)}))\right| = \left|\varphi_n(y_j^{(n)}) - \left(\frac{2}{\pi}\right)^{1/2} \sin\left(j - \frac{1}{2}\right)\pi\right|$$
$$< \frac{1}{3}\left(\frac{2}{\pi}\right)^{1/2}.$$

Hence $\varphi_n(y_j^{(n)}) \neq 0$, and the sign of $\varphi_n(y_j^{(n)})$ is $(-1)^{j-1}$. By the Intermediate Value Theorem and by the fact that φ_n has only $n-1$ zeros in the open interval $0 < x < \beta$, we see that φ_n has one and only one zero $x_j^{(n)}$ in each of the open intervals $(y_j^{(n)}, y_{j+1}^{(n)})$, $j = 1, 2, \cdots, n-1$. Since $g(x)$ is strictly increasing, we have

$$j - \tfrac{1}{2} = ng(y_j^{(n)}) < ng(x_j^{(n)}) < ng(y_{j+1}^{(n)}) = j + \tfrac{1}{2}.$$

Thus when $n \geqq n_1$, for x in the interval $(0, \beta)$, if $N_n(x) = j$, i.e., $x_j^{(n)} \leqq x < x_{j+1}^{(n)}$, then

$$j - \tfrac{1}{2} < ng(x) < j + \tfrac{3}{2}.$$

Now it is clear that for $n \geqq n_1$, $0 \leqq x \leqq \beta$, the following inequality holds:

$$\left|g(x) - \frac{N_n(x)}{n}\right| < \frac{3}{2n}. \tag{2.11}$$

Equation (2.9) follows immediately from (2.11). This completes the proof of Theorem 2.3. □

Now we evaluate the limit (2.3).

THEOREM 2.4.

$$\lim_{n\to\infty}\frac{1}{n^3}\sum_{j=1}^{n}\frac{1}{(x_j^{(n)}-x_{j-1}^{(n)})^2}=\frac{\int_0^\beta[\rho(x)]^{3/2}\,dx}{[\int_0^\beta\sqrt{\rho(x)}\,dx]^3}. \tag{2.12}$$

Proof. By (1.1) we have

$$\begin{aligned}\frac{\lambda_n}{n^2}\cdot\frac{\sum_{j=1}^{n}\min(\rho,I_{n,j})}{n}&\leqq\frac{\pi^2}{n^3}\sum_{j=1}^{n}\frac{1}{(x_j^{(n)}-x_{j-1}^{(n)})^2}\\&\leqq\frac{\lambda_n}{n^2}\cdot\frac{\sum_{j=1}^{n}\max(\rho,I_{n,j})}{n}.\end{aligned} \tag{2.13}$$

Using the formulae $\lim_{n\to\infty}\lambda_n n^{-2}=\pi^2(\int_0^\beta\sqrt{\rho(x)}\,dx)^{-2}$, (2.7), (2.13), and the uniform continuity of $\rho(x)$, we obtain (2.12). □

Remark. Formula (2.7) has the following interesting consequence: "Suppose we can 'hear' infinitely many eigenvalues λ_{n_j} of the string equation (0.1), and we can 'see' the corresponding nodal points $x_1^{(n_j)},\cdots,x_{n_j-1}^{(n_j)}$ of λ_{n_j}, then $\rho(x)$ can be determined." Since by (2.7) and $\lim_{j\to\infty}\lambda_{n_j}n_j^{-2}=\pi^2/(\int_0^\beta\sqrt{\rho(x)}\,dx)^2$, we have all the moments c_n of $\sqrt{\rho(x)}$, where c_n is defined as follows:

$$c_n=\int_0^\beta x^n\sqrt{\rho(x)}\,dx,\qquad n=1,2,\cdots.$$

By the theory of moments (see [4, Chap. III]), $\sqrt{\rho(x)}$ can be determined from the moments c_n.

REFERENCES

[1] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1956.

[2] B. M. LEVITAN AND I. S. SARGSJAN, *Introduction to Spectral Theory*, Translations of Mathematical Monographs, Vol. 39, American Mathematical Society, Providence, RI, 1975.

[3] G. PÓLYA AND G. SZEGÖ, *Problems and Theorems in Analysis*, Vol. I, Die Grundlehren der mathematischen Wissenschaften, Band 193, Springer-Verlag, Berlin, New York, 1972.

[4] D. V. WIDDER, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

014

A RESOLUTION METHOD FOR RICCATI DIFFERENTIAL SYSTEMS COUPLED IN THEIR QUADRATIC TERMS*

L. JODAR† AND H. ABOU-KANDIL†

Abstract. By means of algebraic transformations a Riccati differential matrix system coupled in its quadratic terms is reduced to another one for which the successive approximation method is available. An iterative algorithm for solving the problem and an error upper bound for the approximation are given.

Key words. iterative algorithm, matrix differential system, initial value problem, convergence interval, error bound, approximate solution, coupled Riccati system

AMS(MOS) subject classifications. 34A15, 34A50, 65L05, 65F35, 15A24

1. Introduction. When trying to solve many interesting control problems we are often led to a set of coupled Riccati matrix differential equations. Examples are found in singular [3, p. 41] and hybrid system control [9], reduced order compensator design [2], and nonzero sum differential games [12]. Depending on the problem under consideration, different types of coupling will appear. However, the common feature in all the cases mentioned above is that such differential systems are difficult to solve. A general class of coupled Riccati equations is considered here, i.e.,

$$\begin{aligned} \dot{K}_1 &= -Q_1(t) - A_1(t)K_1 - K_1A_2(t) + K_1S_1(t)K_1 + K_1S_2(t)K_2, \\ \dot{K}_2 &= -Q_2(t) - B_1(t)K_2 - K_2B_2(t) + K_2S_2(t)K_2 + K_2S_1(t)K_1. \end{aligned} \tag{1.1}$$

The problem is then to find $K_1(t)$, $K_2(t)$, for $t \in [0, t_f]$ with the terminal conditions

$$K_1(t_f) = K_{1f}, \qquad K_2(t_f) = K_{2f}. \tag{1.2}$$

It is clear here that the coupling appears only through the quadratic terms. In fact (1.1) is a generalized form for Riccati systems appearing when a Nash equilibrium solution is sought for a two-player linear differential game with quadratic cost functionals [12]. For the time-invariant case and under some assumptions relating the constant coefficients, a number of methods have been developed to obtain numerical [11] or series solutions [5]. Moreover, under the further assumption: $Q_2 = \alpha Q_1$, where α is a scalar, an explicit solution has been proposed in [1].

The purpose of this paper is to show that after an appropriate algebraic transformation, the successive approximation method may be used to solve (1.1). This leads to a straightforward algorithm for which the convergence conditions are clearly stated. Furthermore, a bound of the approximation error is obtained so that the number of iterations required for a given precision can be predetermined. The solution procedure developed here is related to the method proposed in [7] to solve coupled Lyapunov equations.

2. Basic notation and preliminary results. In order to make the paper somewhat self-contained, definitions and results to be used later are recalled in this section.

* Received by the editors December 1, 1986; accepted for publication (in revised form) December 15, 1987.

† Department of Applied Mathematics, Polytechnical University of Valencia, P.O. Box 22.012, Valencia, Spain.

‡ Laboratoire des Signaux et Systèmes, C.N.R.S.-E.S.E., Plateau du Moulon, 91.190 Gif-sur-Yvette, France.

If A and B are matrices in $R_{m\times n}$ and $R_{k\times s}$, respectively, then the tensor product of A and B, denoted $A\otimes B$, is defined as the partitioned matrix

$$A\otimes B=\begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & & & \vdots \\ a_{m1}B & a_{n2}B & \dots & a_{mn}B \end{bmatrix}.$$

If $A\in R_{m\times n}$, we denote

$$A_{.j}=\begin{bmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{bmatrix}, \qquad 1\leqq j\leqq n$$

and

$$\operatorname{vec} A=\begin{bmatrix} A_{.1} \\ \vdots \\ A_{.n} \end{bmatrix}.$$

Note that if $A=[A_1, A_2]$, then

$$\operatorname{vec} A=\begin{bmatrix} \operatorname{vec} A_1 \\ \operatorname{vec} A_2 \end{bmatrix}.$$

If M, N, and P are matrices of suitable dimensions, then using the column lemma [10], we get

$$\operatorname{vec}(MNP)=(P^T\otimes M)\operatorname{vec} N \tag{2.1}$$

where P^T is the transpose of P.

The Frobenius norm of a matrix $A\in R_{m\times n}$ is defined by

$$\|A\|_F=\left\{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right\}^{1/2}. \tag{2.2}$$

In the sequel, since only the Frobenius norm is used, it will be simply denoted by $\|\ \|$. The following properties are then verified for $A\in R_{m\times n}$ and $B\in R_{n\times q}$ [6, p. 274]:

$$\|AB\|\leqq\|A\|\,\|B\|, \tag{2.3}$$

$$\|A\|=\|\operatorname{vec} A\|, \tag{2.4}$$

$$\|A\otimes B\|=\|A\|\,\|B\|. \tag{2.5}$$

Finally, if $F_{ij}\in R_{n\times n}$, for $1\leqq i, j\leqq 2$, then

$$\left\|\begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix}\right\|\leqq 4\max\{\|F_{ij}\|;\ 1\leqq i, j\leqq 2\}. \tag{2.6}$$

The following theorem may now be stated.

THEOREM 1. *Let K_i, P_i be matrices in $R_{n\times n}$ for $i=1, 2$, with $K=[K_1, K_2]$, $P=[P_1, P_2]$, and let $S=[S_1, S_2]$ be a fixed matrix in $R_{n\times 2n}$; then if $\Psi_S: R_{2n}2_{x1}\to R_{2n}2_{x1}$ is the function defined by the expression*

$$\Psi_S(\operatorname{vec} K)=\begin{bmatrix} K_1^T\otimes K_1 & K_2^T\otimes K_1 \\ K_1^T\otimes K_2 & K_2^T\otimes K_2 \end{bmatrix}[\operatorname{vec} S], \tag{2.7}$$

it follows that

$$\|\Psi_S(\text{vec } K)-\Psi_S(\text{vec } P)\| \leqq 4\|S\|(\|K\|+\|P\|)\|\text{vec } K-\text{vec } P\|. \tag{2.8}$$

Proof. From the definition of Ψ_S, we have

$$\Psi_S(\text{vec } K)-\Psi_S(\text{vec } P)=\begin{bmatrix} K_1^T\otimes K_1-P_1^T\otimes P_1 & K_2^T\otimes K_1-P_2^T\otimes P_1 \\ K_1^T\otimes K_2-P_1^T\otimes P_2 & K_2^T\otimes K_2-P_2^T\otimes P_2 \end{bmatrix}[\text{vec } S]. \tag{2.9}$$

Taking norms on both sides and using (2.3), (2.6), we find that

$$\|\Psi_S(\text{vec } K)-\Psi_S(\text{vec } P)\| \leqq 4\|\text{vec } S\|\max\{\|K_i^T\otimes K_j-P_i^T\otimes P_j\|;\ 1\leqq i\, j\leqq 2\}. \tag{2.10}$$

The elements of the partitioned matrix (2.9) can be rewritten as

$$K_i^T\otimes K_j-P_i^T\otimes P_j=(K_i^T-P_i^T)\otimes K_j+P_i^T\otimes(K_j-P_j), \qquad 1\leqq i, j\leqq 2; \tag{2.11}$$

then

$$\|K_i^T\otimes K_j-P_i^T\otimes P_j\| \leqq \|K_i-P_i\|\,\|K_j\|+\|P_i\|\,\|K_j-P_j\|, \tag{2.12}$$

since

$$\|K_i\|=\|\text{vec } K_i\| \leqq \|\text{vec}(K)\|, \qquad \|P_i\|=\|\text{vec } P_i\| \leqq \|\text{vec}(P)\|, \quad i=1,2.$$

Equation (2.12) leads to

$$\|K_i^T\otimes K_j-P_i^T\otimes P_j\| \leqq \|\text{vec } K-\text{vec } P\|(\|\text{vec } K\|+\|\text{vec } P\|). \tag{2.13}$$

Considering (2.10) and (2.13), we immediately obtain the result given in (2.8).

COROLLARY 1. *Let* $M=(M_1, M_2)\in R_{n\times 2n}$ *and* $\delta>0$. *If* K, P *are matrices in* $R_{n\times 2n}$ *such that* $\|K-M\| \leqq \delta$, $\|P-M\| \leqq \delta$, *then*

$$\|\Psi_S(\text{vec } K)-\Psi_S(\text{vec } P)\| \leqq 8\|S\|(\delta+\|M\|)\|\text{vec } K-\text{vec } P\|. \tag{2.14}$$

Proof. It is clear that (2.14) is a direct consequence of (2.8).

3. Main results. The initial value problem (1.1)-(1.2) is now considered. By introducing tensor products in the two members of (1.1), and taking into account the column lemma (equation 2.1) [10], it follows that

$$\begin{aligned} d/dt(\text{vec } K_1(t)) &= -\text{vec } Q_1(t)-(I\otimes A_1(t))(\text{vec } K_1(t))-(A_2^T(t)\otimes I)(\text{vec } K_1(t)) \\ &\quad +(K_1^T(t)\otimes K_1(t))(\text{vec } S_1(t))+(K_2^T(t)\otimes K_1(t))(\text{vec } S_2(t)), \\ d/dt(\text{vec } K_2(t)) &= -\text{vec } Q_2(t)-(I\otimes B_1(t))(\text{vec } K_2(t))-(B_2^T(t)\otimes I)(\text{vec } K_2(t)) \\ &\quad +(K_2^T(t)\otimes K_2(t))(\text{vec } S_2(t))+(K_1^T(t)\otimes K_2(t))(\text{vec } S_1(t)), \\ \text{vec } K_1(t_f) &= \text{vec } K_{1f}, \qquad \text{vec } K_2(t_f)=\text{vec } K_{2f}. \end{aligned} \tag{3.1}$$

Define

$$D_1(t)=(I\otimes A_1(t))+(A_2(t)\otimes I), \qquad D_2(t)=(I\otimes B_1(t))+(B_2(t)\otimes I). \tag{3.2}$$

System (3.1) can be rewritten as

$$\begin{aligned} \frac{d}{dt}(\text{vec } K(t)) &= -\text{vec } Q(t)-\begin{bmatrix} D_1(t) & 0 \\ 0 & D_2(t) \end{bmatrix}[\text{vec } K(t)] \\ &\quad +\begin{bmatrix} K_1^T(t)\otimes K_1(t) & K_2^T(t)\otimes K_1(t) \\ K_1^T(t)\otimes K_2(t) & K_2^T(t)\otimes K_2(t) \end{bmatrix}[\text{vec } S(t)], \\ \text{vec } K(t_f) &= \text{vec}(K_{1f}, K_{2f}) \end{aligned} \tag{3.3}$$

where $Q(t)=[Q_1(t), Q_2(t)]$, $S(t)=[S_1(t), S_2(t)]$, and $K(t)=[K_1(t), K_2(t)]$ are matrices in $R_{n\times 2n}$, for all $t\in[0, t_f]$. Writing (3.3) in a compact form, we have

$$\begin{aligned} &d/dt(\operatorname{vec} K(t)) = F(t, \operatorname{vec} K(t)),\\ &\operatorname{vec} K(t_f) = \operatorname{vec}(K_{1f}, K_{2f}) \end{aligned} \tag{3.4}$$

where $F: [0, t_f]\times R_{2n^2\times 1}\to R_{2n^2\times 1}$, is defined by (3.3); i.e.,

$$F(t, \operatorname{vec} K(t)) = -\operatorname{vec} Q(t) - D(t)(\operatorname{vec} K(t)) + \Psi_{S(t)}(\operatorname{vec} K(t)) \tag{3.5}$$

with

$$D(t)=\begin{bmatrix} D_1(t) & 0\\ 0 & D_2(t)\end{bmatrix}$$

and $\Psi_{S(t)}$ is given by (2.7).

Assuming the continuity of the coefficient matrix functions occurring in (1.1) on the interval $[0, t_f]$, the following constants are finite and well defined:

$$\begin{aligned} &s=\max_{0\leqq t\leqq t_f}\|\operatorname{vec} S(t)\|, \qquad q=\max_{0\leqq t\leqq t_f}\|\operatorname{vec} Q(t)\|,\\ &d=n^{1/2}\max_{0\leqq t\leqq t_f}\{\|A_i(t)\|, \|B_i(t)\|;\ i=1,2\}. \end{aligned} \tag{3.6}$$

When we exploit the special structure of $D(t)$ and use (2.2), it is easy to show that

$$\max_{0\leqq t\leqq t_f}\|D(t)\|\leqq 2\max_{0\leqq t\leqq t_f}\{\|D_i(t)\|, 1\leqq i\leqq 2\}\leqq 4d \tag{3.7}$$

because the Frobenius norm of the identity matrix in $R_{n\times n}$ is $\|I\|=n^{1/2}$, and, from (2.5), we have $\|D_1(t)\|\leqq n^{1/2}(\|A_1(t)\|+\|A_2(t)\|)$ and $\|D_2(t)\|\leqq n^{1/2}(\|B_1(t)\|+\|B_2(t)\|)$.

The following theorem gives a successive approximation procedure used to solve the problem (1.1)-(1.2) and an upper bound for the approximation error.

THEOREM 2. *Let $\delta>0$, $\gamma=\|\operatorname{vec} K_f\|+\delta$, and let M be the constant defined by the expression*

$$M=q+2d\gamma+4s\gamma^2 \tag{3.8}$$

where q, d, and s are given by (3.6). *If $\alpha=\min\{t_f, \delta/M\}$ and assuming that the coefficient matrix functions $A_i(t)$, $B_i(t)$ and $S_i(t)$ are continuous on the interval $[0, t_f]$, then problem* (1.1)-(1.2) *has a unique solution $K(t)$ on the interval $[t_f-\alpha, t_f]$ such that $K(t)$ is the Frobenius norm limit of the sequence of successive approximations*:

$$\{K^{(p)}(t)\}_{p\geqq 0},\ \textit{where } K^{(p)}=[K_1^{(p)}, K_2^{(p)}] \textit{ are given by } K_i^{(0)}(t)=K_{if},\quad i=1,2$$

and

$$\begin{aligned} K_1^{(p+1)}(t) &= K_{1f}-\int_{t_f}^{t}(Q_1(u)+A_1(u)K_1^{(p)}(u)+K_1^{(p)}(u)A_2(u))\,du\\ &\quad+\int_{t_f}^{t}(K_1^{(p)}(u)S_1(u)K_1^{(p)}(u)+K_1^{(p)}(u)S_2(u)K_2^{(p)}(u))\,du,\\ K_2^{(p+1)}(t) &= K_{2f}-\int_{t_f}^{t}(Q_2(u)+B_1(u)K_2^{(p)}(u)+K_2^{(p)}(u)B_2(u))\,du\\ &\quad+\int_{t_f}^{t}(K_2^{(p)}(u)S_2(u)K_2^{(p)}(u)+K_2^{(p)}(u)S_1(u)K_1^{(p)}(u))\,du. \end{aligned} \tag{3.9}$$

Moreover, for $t \in [t_f - \alpha, t_f]$, *the error upper bound for the pth approximation is given by*

$$e_p(t) = \|\text{vec}\,(K(t)) - \text{vec}\,(K^{(p)}(t))\| \leqq \frac{M(\alpha\rho)^{p+1}}{\rho(p+1)!}\exp(\alpha\rho) \tag{3.10}$$

where

$$\rho = 8\gamma s + 2d. \tag{3.11}$$

Proof. Using the definition of $F(t, \text{vec}\,K(t))$ (equation (3.5)) and (3.6), we get

$$\|F(t, \text{vec}\,K)\| \leqq q + 2d\|\text{vec}\,K\| + 4s\|\text{vec}\,K\|^2; \tag{3.12}$$

thus

$$\begin{aligned}\sup\{\|F(t, \text{vec}\,K)\|;\ 0 \leqq t \leqq t_f;\ \|K - K_f\| \leqq \delta\} &\leqq q + 2d(\delta + \|K_f\|) + 4s(\delta + \|K_f\|)^2 \\ &= q + 2d\gamma + 4s\gamma^2.\end{aligned} \tag{3.13}$$

From Theorem 1, we have that $F(t, \text{vec}\,K)$ satisfies a Lipschitz condition of the type

$$\|F(t, \text{vec}\,K) - F(t, \text{vec}\,P)\| \leqq \rho\|\text{vec}\,K - \text{vec}\,P\|$$

where ρ is given by (3.11) and $\|\text{vec}\,K - \text{vec}\,K_f\| \leqq \delta$, $\|\text{vec}\,P - \text{vec}\,K_f\| \leqq \delta$. Now, from the theorem of the successive approximations [8, p. 129], [4, Chap. 5], the unique solution of the problem (3.4) on the interval $[t_f - \alpha, t_f]$, is given by the Frobenius norm limit of the sequence $\{\text{vec}\,K^{(p)}(t)\}_{p \geqq 0}$, where

$$(\text{vec}\,K^{(0)}(t)) = (\text{vec}\,K_f)$$

and

$$\begin{aligned}\text{vec}\,K^{(p+1)}(t) &= \text{vec}\,K_f + \int_{t_f}^{t} F(u, \text{vec}\,K^{(p)}(u))\,du \\ &= \text{vec}\,K_f + \int_{t_f}^{t} \begin{bmatrix} -Q_1(u) - D_1(u)K_1^{(p)}(u) \\ -Q_2(u) - D_2(u)K_1^{(p)}(u) \end{bmatrix} du \\ &+ \int_{t_f}^{t} \begin{bmatrix} (K_1^{(p)}(u) \otimes K_1^{(p)}(u))(\text{vec}\,S_1(u)) + ((K_2^{(p)}(u))^T \otimes K_1^{(p)}(u))(\text{vec}\,S_2(u)) \\ ((K_1^{(p)}(u))^T \otimes K_2^{(p)}(u))(\text{vec}\,S_1(u)) + ((K_2^{(p)}(u))^T \otimes K_2^{(p)}(u))(\text{vec}\,S_2(u)) \end{bmatrix} du.\end{aligned} \tag{3.14}$$

When we take into account the column lemma [10] and the definitions (3.2), the sum of the two integrands occurring on the right-hand side of (3.14) is equivalent to the following expression:

$$\text{vec}\,(W_1^{(p)}(u), W_2^{(p)}(u))$$

where

$$\begin{aligned}W_1^{(p)}(u) &= -Q_1(u) - A_1(u)K_1^{(p)}(u) - K_1^{(p)}(u)A_2(u) \\ &\quad + K_1^{(p)}(u)S_1(u)K_1^{(p)}(u) + K_1^{(p)}(u)S_2(u)K_2^{(p)}(u), \\ W_2^{(p)}(u) &= -Q_2(u) - B_1(u)K_2^{(p)}(u) - K_2^{(p)}(u)B_2(u) \\ &\quad + K_2^{(p)}(u)S_2(u)K_2^{(p)}(u) + K_2^{(p)}(u)S_1(u)K_1^{(p)}(u).\end{aligned} \tag{3.15}$$

Equation (3.9) is directly obtained using (3.14) and (3.15). The error upper bound of the pth approximation defined by (3.10) is a consequence of the theorem of the successive approximations [8, p. 129], [4, Chap. 5].

Remark. It should be noted that the convergence interval for the proposed algorithm described in Theorem 2 depends on the ratio δ/M, where δ is a free positive number, while M is a function of δ as given by (3.8). Therefore, it is interesting to find a $\delta > 0$ such that δ/M is maximized, because the convergence interval is $[t_f - \alpha, t_f]$, and $\alpha = \min\{t_f, \delta/M\}$. In order to maximize δ/M, let us consider

$$r(\delta) = \delta/M = \delta\{q + 2d(\|\text{vec } K_f\| + \delta) + 4s(\|\text{vec } K_f\| + \delta)^2\}^{-1}.$$

If we assume that all elements of $Q(t) = [Q_1(t), Q_2(t)]$, $S(t) = [S_1(t), S_2(t)]$, are not identically zero, then maximum of $r(\delta)$ is obtained when $d/d\delta(r(\delta)) = 0$, i.e., for

$$\delta^* = \{(4s)^{-1}(q + 2d\|\text{vec } K_f\| + 4s\|\text{vec } K_f\|^2)\}^{1/2}.$$

Acknowledgment. The authors thank an anonymous referee for his constructive remarks.

REFERENCES

[1] H. ABOU-KANDIL AND P. BERTRAND, *Analytic solution for a class of linear quadratic open-loop Nash games*, Internat. J. Control, 43 (1986), pp. 997-1002.

[2] D. S. BERNSTEIN AND D. C. HYLAND, *The optimal projection equations for finite dimensional fixed order dynamic compensation of infinite dimension systems*, SIAM J. Control Optim., 24 (1986), pp. 122-151.

[3] S. L. CAMPBELL, *Singular Systems of Differential Equations* II, Pitman, San Francisco, 1982.

[4] E. A. CODDINGTON, *An Introduction to Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[5] J. B. CRUZ JR. AND C. I. CHEN, *Series Nash solution of two-persons zero-sum linear quadratic games.* J. Optim. Theory Appl., 7 (1971), pp. 240-257.

[6] B. P. DEMIDOVICH AND I. A. MARON, *Calculo Numerico Fundamental*, Paraninfo, Madrid, 1977. (In Spanish.)

[7] L. JÓDAR AND M. MARITON, *Explicit solutions for a system of coupled Lyapunov matrix differential equations*, Proc. Edinburgh Math. Soc., 30 (1987), pp. 427-434.

[8] G. E. LADAS AND V. LAKSMIKANTHAM, *Differential Equations in Abstract Spaces*, Academic Press, New York, 1972.

[9] M. MARITON AND P. BERTRAND, *Nonswitching control strategies for continuous-time jump linear quadratic systems*, in Proc. 24th IEEE Conference Decision and Control, Fort Lauderdale, FL, December 11-13, pp. 916-921.

[10] W. E. ROTH, *On direct product matrices*, Bull. Amer. Math. Soc., 40 (1934), pp. 461-468.

[11] M. SIMAAN AND J. B. CRUZ JR., *On the solution of open-loop Nash-Riccati equations in linear-quadratic differential games*, Internat. J. Control, 18 (1973), pp. 57-63.

[12] A. W. STARR AND Y. C. HO, *Non-zero sum differential games*, J. Optim. Theory Appl., 3 (1969), pp. 179-197.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

015

ON SMOOTHEST INTERPOLANTS*

A. PINKUS†

Abstract. This paper is concerned with the problem of characterizing those functions of minimum L^p-norm on their nth derivative, $1 \leqq p \leqq \infty$, that sequentially take on the given values $(e_i)_1^N$. For $p = \infty$ the unique minimizing function is characterized. For $p < \infty$ fairly explicit necessary conditions are given.

Key words. minimum norm interpolation, Sobolev spaces, splines

AMS(MOS) subject classifications. 41A05, 46E35

1. Introduction. Let $n \geqq 2$ be fixed. For $p \in (1, \infty]$, $W_p^{(n)}$ will denote the usual Sobolev space of real-valued functions on $[0, 1]$ with $n-1$ absolutely continuous derivatives and nth derivative existing almost everywhere as a function in $L^p[0, 1]$. Equivalently,

$$W_p^{(n)} = \left\{ f: f(x) = \sum_{i=0}^{n-1} a_i x^i + \frac{1}{(n-1)!} \int_0^1 (x-y)_+^{n-1} h(y)\, dy,\ h \in L^p,\ a_i \in R,\ i = 0, 1, \cdots, n-1 \right\}.$$

(Here $a_i = f^{(i)}(0)/i!$, $i = 0, 1, \cdots, n-1$, and $h \equiv f^{(n)}$.) For $p = 1$, rather than considering the analogous $W_1^{(n)}$, we introduce $V^{(n)}$. To define $V^{(n)}$, let M denote the space of real Baire measures on $[0, 1]$. For $\mu \in M$, $\|\mu\|$ will denote the total variation of the measure μ. Then

$$V^{(n)} = \left\{ f: f(x) = \sum_{i=0}^{n-1} a_i x^i + \frac{1}{(n-1)!} \int_0^1 (x-y)_+^{n-1}\, d\mu(y),\ \|\mu\| < \infty \right\}.$$

We will shortly explain our reasons for considering $V^{(n)}$ rather than $W_1^{(n)}$.

Let $e_1, \cdots, e_N$ be given real fixed data, $e_i \neq e_{i+1}$, $i = 1, \cdots, N-1$. Set

$$\Xi_N = \{\mathbf{t}: \mathbf{t} = (t_1, \cdots, t_N),\ 0 \leqq t_1 < \cdots < t_N \leqq 1\}.$$

For each $\mathbf{t} \in \Xi_N$, set

$$W_p^{(n)}(\mathbf{t}; \mathbf{e}) = \{f: f \in W_p^{(n)}, f(t_i) = e_i,\ i = 1, \cdots, N\}$$

for $p \in (1, \infty]$, and

$$V^{(n)}(\mathbf{t}; \mathbf{e}) = \{f: f \in V^{(n)}, f(t_i) = e_i,\ i = 1, \cdots, N\}.$$

The following problems are considered in [2] and [6]:

$$\inf \{\|f^{(n)}\|_p : f \in W_p^{(n)}(\mathbf{t}; \mathbf{e})\} \tag{1}$$

for $p \in (1, \infty]$, and

$$\inf \{\|\mu\| : f \in V^{(n)}(\mathbf{t}; \mathbf{e})\}. \tag{2}$$

(It is understood that f and μ in (2) are related as in the definition of $V^{(n)}$.) Later we will describe the solutions to problems (1) and (2). An understanding of their exact form is crucial to a solution of the problems we consider. We do note, however, that there exist $f \in W_p^{(n)}(\mathbf{t}; \mathbf{e})$ and $f \in V^{(n)}(\mathbf{t}; \mathbf{e})$ for which the above infima are in fact attained.

* Received by the editors October 22, 1986; accepted for publication (in revised form) December 15, 1987.
† Technion, Israel Institute of Technology, Haifa 32000, Israel.

If we replace the extremum problem (2) by the analogous problem where $W_1^{(n)}$ takes the place of $V^{(n)}$, then this is not necessarily the case, i.e., the infimum need not be attained by some $f \in W_1^{(n)}$. However, the value of the infimum in (2), or in (2) with $W_1^{(n)}$ replacing $V^{(n)}$, is the same. This is one reason for considering $V^{(n)}$ rather than $W_1^{(n)}$. A more detailed discussion of this matter can be found in de Boor [2], and in Fisher and Jerome [5], [6].

In this work, we are interested in solutions to the problems

$$\inf_{\mathbf{t}\in\Xi_N} \inf\{\|f^{(n)}\|_p : f \in W_p^{(n)}(\mathbf{t}; \mathbf{e})\} \tag{3}$$

for $p \in (1, \infty]$, and

$$\inf_{\mathbf{t}\in\Xi_N} \inf\{\|\mu\| : f \in V^{(n)}(\mathbf{t}; \mathbf{e})\}, \tag{4}$$

and in functions for which the infima are attained. Thus, for each fixed $n \geqq 2$, $p \in [1, \infty]$, and data $(e_i)_1^N$, we wish to characterize those functions f that take on the values $(e_i)_1^N$ sequentially, and minimize the L^p-norm of their nth derivative. (Here we are abusing notation in the case where $p = 1$.)

Before continuing, we note three simple facts.

(I) It suffices to assume that $(e_i - e_{i-1})(e_{i+1} - e_i) < 0$, $i = 2, \cdots, N-1$. This follows from continuity considerations. If, for example, $e_{i-1} < e_i < e_{i+1}$ for some $i \in \{2, \cdots, n-1\}$, then we may delete the condition $f(t_i) = e_i$ since f will always attain the value e_i at some point in (t_{i-1}, t_{i+1}).

(II) We may assume that $N > n$. If $N \leqq n$, then for any choice of $\mathbf{t} \in \Xi_N$, there exists a polynomial q of degree $\leqq n-1$ for which $q(t_i) = e_i$, $i = 1, \cdots, N$. Moreover $q^{(n)} \equiv 0$ and our problem is trivially solved.

(III) We always have $t_1 = 0$ and $t_N = 1$. Assume, for example, that $t_N < 1$ for some f which solves (3) or (4). Set $g(x) = f(xt_N)$ for $x \in [0, 1]$. Then g is "admissible" in (3) or (4), and since $g^{(n)}(x) = t_N^n f^{(n)}(xt_N)$, it easily follows that $\|g^{(n)}\|_p < \|f^{(n)}\|_p$ for $p \in (1, \infty]$, with the analogous strict inequality in (4).

Thus in what follows we will always assume that

(a) $(e_i - e_{i-1})(e_{i+1} - e_i) < 0$, $i = 2, \cdots, N-1$;

(b) $N > n$;

(c) $t_1 = 0$, $t_N = 1$.

There always exist functions $f \in W_p^{(n)}$ which solve (3) (or $f \in V^{(n)}$ which solve (4)). The proof of this fact is not difficult and we omit it. It follows from the existence already alluded to in (1) and (2), and from the fact that there exists a $\mathbf{t}^* \in \Xi_N$ (and not in $\bar{\Xi}_N \backslash \Xi_N$) for which the left-most infima in (3) or (4) are attained.

We will prove that solutions to (3) and (4) must be of a particular form, given by solutions to (1) and (2), respectively, and must also "oscillate" strictly between the values $(e_i)_1^N$. To explain what we mean by this latter term, we introduce the following definition.

DEFINITION. Let $(e_i - e_{i-1})(e_{i+1} - e_i) < 0$, $i = 2, \cdots, N-1$, and $0 = t_1 < \cdots < t_N = 1$. Let $f \in C[0, 1]$ satisfy $f(t_i) = e_i$, $i = 1, \cdots, N$. We say that f *oscillates* between the $(e_i)_1^N$ on $(t_i)_1^N$ if f is monotone on $[t_i, t_{i+1}]$ for each $i = 1, \cdots, N-1$. We say that f *oscillates strictly* between the $(e_i)_1^N$ on $(t_i)_1^N$ if f is strictly monotone on $[t_i, t_{i+1}]$ for each $i = 1, \cdots, N-1$.

Because of the nature of the problem, we divide our analysis into three parts, namely, $1 < p < \infty$, $p = \infty$, and $p = 1$. Both $p = 1 (V^{(n)})$ and $p = \infty$ may be considered as limiting cases, but they are much more special and will be considered separately.

In § 2 we quite easily prove that solutions to (3) for $p \in (1, \infty)$ must be of a particular form and oscillate strictly between the $(e_i)_1^N$, on some $(t_i)_1^N$. However, we are unable to prove either the uniqueness of the solution or the fact that functions of this particular form are necessarily solutions to (3). In other words, we prove necessary but not sufficient conditions for a solution to (3). We do conjecture, however, that these conditions are sufficient and that (3) has a unique solution.

In § 3 we consider the case $p = \infty$. We somewhat surprisingly are able to explicitly characterize the solution and we show that it is unique. The uniqueness is especially surprising since for fixed $\mathbf{t} \in \Xi_N$ and $p = \infty$, the solution to (1) is not necessarily unique. (While for $p \in (1, \infty)$ the solution to (1) is unique.)

In § 4 we consider the case $p = 1$. We prove that the solution to (2) is unique and is of a particularly simple form (splines of degree $n-1$ with $N-n$ knots). We again prove a necessary condition for the solution to (4). Here again both the full characterization and uniqueness is lacking, except in the case $n = 2$ where it is easily seen that every solution to (2) necessarily oscillates strictly between the $(e_i)_1^N$. For $n \geqq 3$, we conjecture that the characterization leads to a unique solution.

Let us review the history of and motivation behind this problem. The above problem in a multidimensional setting (x still runs over $[0, 1]$, but we are dealing with a d-dimensional vector of single-valued functions and d-dimensional data vectors $\mathbf{e}^i$, $i = 1, \cdots, N$) was discussed by Marin [9] and Töpfer [12]. Physical motivation for this problem comes from problems of geometric curve fitting and design of a trajectory for a robot manipulator (see Marin [9] and Töpfer [12]). Marin explicitly proved existence and uniqueness for the one-dimensional problem in the case $p = 2$ and $n = 2$. Here we are dealing with natural cubic splines (solutions of (1)) and we can explicitly calculate the solution. More recently Scherer and Smith [11] dealt with the problem of existence in the multidimensional setting for the case $p = 2$. It is our hope that the one-dimensional problem considered herein will not only be of interest in and of itself but will also provide insight into the multidimensional problem.

Finally it should be noted that generalizations of the results of this paper exist since many of these results are consequences of the underlying total positivity structure of the problem. However, this is not true of all of the results and generally only weaker versions hold. Thus, for example, we might consider an nth order disconjugate differential equation L on $[0, 1]$, and the problem

$$\inf_{\mathbf{t} \in \Xi_N} \inf \{ \| Lf \|_p : f \in W_p^{(n)}(\mathbf{t}; \mathbf{e}) \}$$

for $p \in (1, \infty]$, with an analogue of (4) for $p = 1$. The main result of § 2, Theorem 2.2, will hold in an analogous form except that it is not necessary that $t_1 = 0$, $t_N = 1$, or that every optimal f^* oscillate strictly between the $(e_i)_1^N$ on some $(t_i^*)_1^N$, but only that f^* on $[t_i^*, t_{i+1}^*]$ take on only values between e_i and e_{i+1}, $i = 1, \cdots, N-1$. For $p = 1$, and especially $p = \infty$, the results are substantially weaker than those obtained herein.

Another generalization that can be dealt with using the techniques of this paper is the following. Consider the problem

$$\inf_{\mathbf{t} \in \Xi_N} \inf \left\{ \| h \|_p : \int_0^1 K(t_i, y) h(y)\, dy = e_i, i = 1, \cdots, N \right\}$$

(and the analogous problem for $p = 1$) where K is a strictly totally positive kernel. Here again weaker results of the above form are obtained, but only in the case where $e_i e_{i+1} < 0$, $i = 1, \cdots, N-1$.

2. $p\in(1,\infty)$. To understand the solution to (3) we must first consider the problem (1).

Recall that for $f\in W_p^{(n)}$,

$$f(x)=\sum_{i=0}^{n-1}a_ix^i+\frac{1}{(n-1)!}\int_0^1(x-y)_+^{n-1}h(y)\,dy, \tag{5}$$

where $a_i=f^{(i)}(0)/i!$, $i=0,1,\cdots,n-1$, and $h\equiv f^{(n)}\in L^p$.

Let $\mathbf{t}\in\Xi_N$ be fixed, $t_1=0$, $t_N=1$, $N\geqq n+1$. $f[t_i,\cdots,t_{i+n}]$ will denote the nth divided difference of f at the points $t_i,\cdots,t_{i+n}$, $i=1,\cdots,N-n$. For $f\in W_p^{(n)}(\mathbf{t};\mathbf{e})$, set

$$E_i=f[t_i,t_{i+1},\cdots,t_{i+n}],\qquad i=1,\cdots,N-n.$$

When we assume $(e_i-e_{i-1})(e_{i+1}-e_i)<0$, $i=2,\cdots,N-1$, it easily follows that $E_iE_{i+1}<0$, $i=1,\cdots,N-n-1$ (since $n\geqq 1$). Applying the nth divided difference at the points $t_i,\cdots,t_{i+n}$ to $f\in W_p^{(n)}(\mathbf{t};\mathbf{e})$ as in (5), we obtain

$$E_i=\int_0^1 M_{i,n}(y)h(y)\,dy,\qquad i=1,\cdots,N-n,$$

where $M_{i,n}$ is a positive multiple (easily computed) of the B-spline of degree $n-1$ with knots $t_i,\cdots,t_{i+n}$, $i=1,\cdots,N-n$. Problem (1) is equivalent to

$$\inf\left\{\|h\|_p\colon\int_0^1 M_{i,n}(y)h(y)\,dy=E_i,\ i=1,\cdots,N-n\right\}. \tag{6}$$

Problem (6) (see de Boor [2], and Fisher and Jerome [6]) (and thus (1)) has a unique solution of the form

$$h_p(y)=\left|\sum_{i=1}^{N-n}b_iM_{i,n}(y)\right|^{q-1}\operatorname{sgn}\left(\sum_{i=1}^{N-n}b_iM_{i,n}(y)\right) \tag{7}$$

where $1/p+1/q=1$, and

$$E_i=\int_0^1 M_{i,n}(y)h_p(y)\,dy,\qquad i=1,\cdots,N-n. \tag{8}$$

Equation (8) uniquely determines the coefficients $(b_i)_1^{N-n}$ in (7). To obtain the unique solution f_p to (1), we write

$$f_p(x)=\sum_{i=0}^{n-1}a_ix^i+\frac{1}{(n-1)!}\int_0^1(x-y)_+^{n-1}h_p(y)\,dy$$

and uniquely determine the $(a_i)_0^{n-1}$ so that $f_p(t_i)=e_i$, $i=1,\cdots,n$. From (8) it follows that $f_p\in W_p^{(n)}(\mathbf{t};\mathbf{e})$.

The following notation will prove useful. For $f\in C[a,b]$, let $S(f)$ denote the number of sign changes of f on $[a,b]$, i.e.,

$$S(f)=\sup\{k\colon a\leqq x_1<\cdots<x_{k+1}\leqq b,\ f(x_i)f(x_{i+1})<0,\ i=1,\cdots,k\}.$$

Of course, if f is either nonnegative or nonpositive on $[a,b]$, then we set $S(f)=0$. Similarly, for a vector $\mathbf{x}\in R^m\backslash\{\mathbf{0}\}$, $S^-(\mathbf{x})$ will denote the number of sign changes of the vector $\mathbf{x}$, i.e.,

$$S^-(\mathbf{x})=\max\{k\colon 1\leqq i_1<\cdots<i_{k+1}\leqq m,\ x_{i_j}x_{i_{j+1}}<0,\ j=1,\cdots,k\},$$

unless $\mathbf{x}$ is nonnegative or nonpositive in which case $S^-(\mathbf{x})=0$.

Let $Q(x)=\sum_{i=1}^{N-n} b_i M_{i,n}(x)$ where the $(b_i)_1^{N-n}$ are as determined by (8).

PROPOSITION 2.1. *Q has exactly $N-n-1$ sign changes on $(0,1)$, and $\sigma Q^{(n-1)}(x)(-1)^i>0$ for $x\in(t_i, t_{i+1})$, $i=1,\cdots,N-1$, where $\sigma\in\{-1,1\}$, fixed.*

Proof. It is well known (see, e.g., de Boor [3]) that

$$S\left(\sum_{i=1}^{N-n} b_i M_{i,n}\right)\leqq S^-((b_1,\cdots,b_{N-n})).$$

Since $S^-((b_1,\cdots,b_{N-n}))\leqq N-n-1$, it follows that Q has at most $N-n-1$ sign changes on $(0,1)$. Assume Q has k sign changes on $(0,1)$. Then

$$h_p(y)=|Q(y)|^{q-1}\operatorname{sgn}(Q(y))$$

has k sign changes on $(0,1)$. Let $0=\xi_0<\xi_1<\cdots<\xi_{k+1}=1$ be such that $\delta h_p(y)(-1)^j\geqq 0$ for all $y\in[\xi_{j-1},\xi_j]$, $j=1,\cdots,k+1$, where $\delta\in\{-1,1\}$, fixed. Set

$$a_{ij}=\int_{\xi_{j-1}}^{\xi_j} M_{i,n}(y)|h_p(y)|\,dy,\qquad i=1,\cdots,N-n,\quad j=1,\cdots,k+1.$$

From properties of B-splines (see, e.g., de Boor [3]) it follows that $A=(a_{ij})_{i=1,j=1}^{N-n,k+1}$ is a totally positive (TP) matrix. Furthermore,

$$\sum_{j=1}^{k+1} a_{ij}(-1)^j\delta=E_i,\qquad i=1,\cdots,N-n.$$

From the above, $N-n\geqq k+1$. As a consequence of the variation diminishing property of TP matrices (see Karlin [7]), we have

$$S^-((E_1,\cdots,E_{N-n}))\leqq\min\{\operatorname{rank}(A)-1, S^-(-\delta,\delta,\cdots,(-1)^{k+1}\delta)\}.$$

Since $E_iE_{i+1}<0$, $i=1,\cdots,N-n-1$, it follows that the left-hand side equals $N-n-1$ and that $N-n-1\leqq k$. Therefore $k=N-n-1$ and Q has exactly $N-n-1$ sign changes on $(0,1)$.

Since $k=N-n-1$, we have that $S^-((b_1,\cdots,b_{N-n}))=N-n-1$, and thus $b_i(-1)^i\sigma>0$, $i=1,\cdots,N-n$, for some $\sigma\in\{-1,1\}$, fixed. It is well known that the $(n-1)$st derivative of $M_{i,n}(x)$ strictly alternates in sign as we go from (t_j,t_{j+1}) to (t_{j+1},t_{j+2}), $j=1,\cdots,i+n-2$. In particular,

$$M_{i,n}^{(n-1)}(x)(-1)^{i+j}>0,\quad x\in(t_j,t_{j+1}),\quad j=i,\cdots,i+n-1.$$

Thus for $x\in(t_j,t_{j+1})$,

$$Q^{(n-1)}(x)=\sum_{i=1}^{N-n} b_iM_{i,n}^{(n-1)}(x)=\sigma(-1)^j\sum_{i=1}^{N-n}|b_iM_{i,n}^{(n-1)}(x)|.$$

Since $b_i\neq 0$ for all i, and $M_{i,n}^{(n-1)}(x)\not\equiv 0$ on (t_j,t_{j+1}) for some i, it follows that $\sigma(-1)^jQ^{(n-1)}(x)>0$ on (t_j,t_{j+1}), $j=1,\cdots,N-1$. This proves the proposition. □

With the above proposition we easily prove the following theorem.

THEOREM 2.2. *Let $p\in(1,\infty)$, and let $f^*\in W_p^{(n)}$ be a solution of* (3). *There exists a $\mathbf{t}^*=(t_1^*,\cdots,t_N^*)$, $0=t_1^*<\cdots<t_N^*=1$ such that $f^*\in W_p^{(n)}(\mathbf{t}^*;\mathbf{e})$. Furthermore,*

$$\text{(a)}\quad f^{*(n)}(y)=\left|\sum_{i=1}^{N-n} b_i^*M_{i,n}(y)\right|^{q-1}\operatorname{sgn}\left(\sum_{i=1}^{N-n} b_i^*M_{i,n}(y)\right)$$

where $1/p+1/q=1$, $M_{i,n}$ is a positive multiple of the B*-spline of degree $n-1$ with knots*

$t_i^*, \cdots, t_{i+n}^*$, $i=1, \cdots, N-n$, *and the* $(b_i^*)_1^{N-n}$ *satisfy*

$$\int_0^1 M_{i,n}(y) f^{*(n)}(y)\, dy = f[t_i^*, \cdots, t_{i+n}^*], \qquad i=1, \cdots, N-n.$$

(b) f^* *oscillates strictly between the* $(e_i)_1^N$ *on* $(t_i^*)_1^N$.

Proof. Let f^* solve (3). Existence implies that there exists a $\mathbf{t}^*$ as above for which $f^* \in W_p^{(n)}(\mathbf{t}^*; \mathbf{e})$. Since f^* must also solve (1) for $\mathbf{t}^*$, it follows that f^* necessarily satisfies (a). It remains to prove that (b) holds.

Since $(e_i - e_{i-1})(e_{i+1} - e_i) < 0$, $i=2, \cdots, N-1$, f^* has at least $N-2$ interior extrema, i.e., $f^{*\prime}$ has at least $N-2$ distinct zeros. From Proposition 2.1, $f^{*(n)}$ does not vanish on any subinterval of $[0, 1]$ and has exactly $N-n-1$ sign changes. From Rolle's theorem applied to $f^{*\prime}$, it follows that $f^{*\prime}$ has exactly $N-2$ (simple) zeros in $(0, 1)$. Let $0 < s_2 < \cdots < s_{N-1} < 1 (s_1 = 0, s_N = 1)$ denote the unique extrema of f^*. Thus f^* oscillates strictly between the $(f^*(s_i))_1^N$ on $(s_i)_1^N$. It remains to prove that $s_i = t_i^*$, $i=2, \cdots, N-1$. Note that $t_{j-1}^* < s_j < t_{j+1}^*$ for $j=2, \cdots, N-1$.

Assume $s_i \neq t_i^*$ for some $i \in \{2, \cdots, N-1\}$. Consider problem (1) at the points $(s_i)_1^N$ with the values $e_i^1 = f^*(s_i)$, $i=1, \cdots, N$. There is a unique solution to this new problem which we denote by g^*. It follows from Proposition 2.1 that $g^* \not\equiv f^*$. Furthermore f^* is "admissible" for this problem. Thus $\|g^{(n)}\|_p < \|f^{*(n)}\|_p$. From continuity considerations, there exist points $0 \leqq w_1 < \cdots < w_N \leqq 1$ such that $g^*(w_i) = e_i$, $i = 1, \cdots, N$. Thus g^* is "admissible" in (3). However, this contradicts the minimality property of f^*. Thus $s_i = t_i^*$, $i=2, \cdots, N-1$, and f^* oscillates strictly between the $(e_i)_1^N$ on $(t_i^*)_1^N$. □

On the basis of the above result, it is natural to ask whether the solution to (3) is unique, and in particular, whether there is a unique function satisfying (a) and (b) of Theorem 2.2. Marin [9] showed by construction that there is a unique function satisfying (a) and (b) in the particular case $n=p=2$.

Remark. For the case $n=1$, it is easily seen that every solution to (1) a spline of degree one with simple knots at $t_2, \cdots, t_{N-1}$. Thus it oscillates strictly between the $(e_i)_1^N$ on $(t_i)_1^N$ for any choice of $\mathbf{t} \in \Xi_N$. However, a bit of calculation shows that the solution to (3) is in fact unique. The optimal choice of $\mathbf{t}^*$ is given by $t_1^* = 0$, $t_N^* = 1$, and

$$t_i^* = \sum_{j=1}^{i-1} |e_{j+1} - e_j| \Big/ \sum_{j=1}^{N-1} |e_{j+1} - e_j|, \qquad i=2, \cdots, N-1.$$

Note that this unique choice is independent of $p \in (1, \infty)$.

3. $p=\infty$. For fixed $0 = t_1 < \cdots < t_N = 1$, $N > n \geqq 2$, and $(e_i - e_{i-1})(e_{i+1} - e_i) < 0$, $i=2, \cdots, N-1$, the problem (1) for $p=\infty$, i.e.,

$$\inf \{\|f^{(n)}\|_\infty : f \in W_\infty^{(n)}(\mathbf{t}; \mathbf{e})\}, \tag{9}$$

may have many solutions. There is always at least one solution of particular interest. It is a perfect spline of degree n with exactly $N-n-1$ knots, i.e., a function P of the form

$$P(x) = \sum_{i=0}^{n-1} a_i x^i + \frac{c}{n!}\left[x^n + 2\sum_{i=1}^{N-n-1} (-1)^i (x - \xi_i)_+^n\right]$$

where $\xi_0 = 0 < \xi_1 < \cdots < \xi_{N-n-1} < \xi_{n-n} = 1$ (see, e.g., Karlin [8]). Note that $|P^{(n)}(x)| = |c|$ for all $x \in [0, 1] \backslash \{\xi_1, \cdots, \xi_{N-n-1}\}$.

The main idea used in the proof of Theorem 2.2 does not carry over to the case $p=\infty$ since the g^* constructed therein is generally identically equal to the f^*. However, much research has been done on perfect splines and we will use some of those results to prove not only an analogue of Theorem 2.2, but also the uniqueness of our solution.

We first state two deep results due to Bojanov. Recall that we always assume that $N > n \geqq 2$ and $(e_i - e_{i-1})(e_{i+1} - e_i) < 0$, $i = 2, \cdots, N-1$.

THEOREM 3.1 (Bojanov [1]). *There exists a unique perfect spline P^* of degree n with $N-n-1$ knots, and a unique set of points $0 = t_1^* < \cdots < t_N^* = 1$ for which*

(i) $P^*(t_i^*) = e_i$, $i = 1, \cdots, N$;

(ii) $P^{*\prime}(t_i^*) = 0$, $i = 2, \cdots, N-1$.

THEOREM 3.2 (Bojanov [1]). *Let $P^*(\cdot\,; \mathbf{e})$ denote the unique perfect spline as given in Theorem 3.1, where we indicate the dependence on $\mathbf{e} = (e_1, \cdots, e_N)$. In a neighborhood of every $\mathbf{e}$ satisfying $(e_i - e_{i-1})(-1)^i > 0$, $i = 2, \cdots, N$, we have that $\|P^{*(n)}(\cdot\,; \mathbf{e})\|_\infty$ is a strictly increasing function of each $(-1)^i e_i$, $i = 1, \cdots, N$.*

On the basis of Theorems 3.1 and 3.2, we immediately obtain the following result.

THEOREM 3.3. *There exists a unique perfect spline P^* of degree n with $N-n-1$ knots which satisfies (3) for $p = \infty$. P^* is uniquely characterized by the fact that it oscillates strictly between the $(e_i)_1^N$ on some $(t_i^*)_1^N$.*

Proof. The only fact that is not an immediate consequence of Theorems 3.1 and 3.2 is the fact that P^* is strictly monotone on $[t_i^*, t_{i+1}^*]$ for each $i = 1, \cdots, N-1$. However, a simple Rolle's theorem argument shows that $P^{*\prime}$ has exactly $N-2$ zeros. Thus P^* is strictly monotone on $[t_i^*, t_{i+1}^*]$, $i = 1, \cdots, N-1$. □

In the above theorem, uniqueness is proved only for the class of perfect splines. There is more that is true, namely, Theorem 3.4.

THEOREM 3.4. *The perfect spline of Theorem 3.3 is the unique solution to (3) for $p = \infty$.*

Proof. Let P^* be as in Theorem 3.3 with $P^*(t_i^*) = e_i$, $i = 1, \cdots, N$, $0 = t_1^* < \cdots < t_N^* = 1$. Assume $f \in W_\infty^{(n)}(\mathbf{t}; \mathbf{e})$ for some $\mathbf{t} \in \Xi_N$. There exists a perfect spline P of degree n with $N-n-1$ knots for which $P(t_i) = f(t_i) = e_i$, $i = 1, \cdots, N$, and $\|P^{(n)}\|_\infty \leqq \|f^{(n)}\|_\infty$. If $\mathbf{t} \neq \mathbf{t}^*$, then from Theorem 3.3, $\|P^{*(n)}\|_\infty < \|P^{(n)}\|_\infty$. Thus if $f \in W_\infty^{(n)}$, f is "admissible" in (3), and $\|f^{(n)}\|_\infty = \|P^{*(n)}\|_\infty$, then it necessarily follows that $f \in W_\infty^{(n)}(\mathbf{t}^*; \mathbf{e})$.

We next prove that $f'(t_i^*) = 0$, $i = 2, \cdots, N-1$ for any f as above. Assume $f'(t_j^*) \neq 0$ for some $j \in \{2, \cdots, N-1\}$. Replace t_j^* by $s_j \in (t_{j-1}^*, t_{j+1}^*)$ so that if $g \in W_\infty^{(n)}$, $g(t_i^*) = e_i$, $i = 1, \cdots, N$, $i \neq j$, and $g(s_j) = f(s_j)$, then g attains the value e_j at least twice in (t_{j-1}^*, t_{j+1}^*), and g is "admissible" in (3). Let P be a perfect spline of degree n with $N-n-1$ knots such that $P(t_i^*) = e_i$, $i = 1, \cdots, N$, $i \neq j$, and $P(s_j) = f(s_j)$. Then $P \not\equiv P^*$. Thus $\|P^{(n)}\|_\infty \leqq \|f^{(n)}\|_\infty$ and from Theorem 3.3, $\|P^{*(n)}\|_\infty < \|P^{(n)}\|_\infty$. This contradicts the minimality property of f. Thus $f'(t_i^*) = 0$, $i = 2, \cdots, N-1$.

Assume $f \not\equiv P^*$. Then $f \not\equiv P^*$ on (t_j^*, t_{j+1}^*) for some $j \in \{1, \cdots, N-1\}$. Since $(P^* - f)(t_i^*) = 0$, $i = j, j+1$, $(P^* - f)'(x)$ must change sign on (t_j^*, t_{j+1}^*). Thus for $\sigma > 0$, sufficiently small, $(P^* - (1-\sigma)f)'(x)$ has a sign change in (t_j^*, t_{j+1}^*). Furthermore $(P^* - (1-\sigma)f)'(t_i^*) = 0$, $i = 2, \cdots, N-1$. Thus $(P^* - (1-\sigma)f)'(x)$ has at least $N-1$ distinct zeros in $[0, 1]$. Since $|P^{*(n)}(x)| \geqq |f^{(n)}(x)| > (1-\sigma)|f^{(n)}(x)|$ almost everywhere on $[0, 1]$, it follows from Rolle's theorem that $(P^* - (1-\sigma)f)^{(n)}(x)$ has at least $N-n$ sign changes on $[0, 1]$. But $P^{*(n)}(x)$, and thus $(P^* - (1-\sigma)f)^{(n)}(x)$, has exactly $N-n-1$ sign changes thereon. This contradiction proves the theorem. □

Remark. For $N = n+1$, P^* is the unique polynomial of degree n that satisfies $P^*(t_i^*) = e_i$, $i = 1, \cdots, n+1$, and $P^{*\prime}(t_i^*) = 0$, $i = 2, \cdots, n$. Such polynomials have been considered previously (see, e.g., Davis [4] and Mycielski and Paszkowski [10]). It seems that it was not previously noted that such polynomials satisfy an extremal property with respect to their nth derivative.

Remark. In the case $n = 1$ both Theorems 3.3 and 3.4 are valid. However, the proofs are somewhat different. The unique solution is identical for all $p \in (1, \infty]$ (see the remark at the end of § 2), and is a perfect spline with knots $(t_i^*)_2^{N-1}$.

4. $p=1$. We first consider in some detail solutions to (2) for fixed $0=t_1<t_2<\cdots<t_N=1$ with $N>n\geqq 2$ and $(e_i-e_{i-1})(e_{i+1}-e_i)<0$, $i=2,\cdots,N-1$. We recall that $f\in V^{(n)}(\mathbf{t};\mathbf{e})$ if

$$f(x)=\sum_{i=0}^{n-1}a_ix^i+\frac{1}{(n-1)!}\int_0^1(x-y)_+^{n-1}\,d\mu(y), \tag{10}$$

where $\|\mu\|<\infty$, and $f(t_i)=e_i$, $i=1,\cdots,N$. We are concerned with the problem

$$\min\{\|\mu\|: f\in V^{(n)}(\mathbf{t};\mathbf{e})\} \tag{11}$$

where μ is associated with f as in (10).

As in § 2, set

$$E_i=f[t_i,t_{i+1},\cdots,t_{i+n}],\qquad i=1,\cdots,N-n.$$

Thus (11) is equivalent to

$$\min\left\{\|\mu\|:\int_0^1 M_{i,n}(y)\,d\mu(y)=E_i,\ i=1,\cdots,N-n\right\}. \tag{12}$$

Since $n\geqq 1$, we have $E_iE_{i+1}<0$, $i=1,\cdots,N-n-1$.

It is well known that $(M_{i,n})_1^{N-n}$ is a weak Chebyshev (WT-) system on $[0,1]$. Thus there exists a nontrivial

$$h(y)=\sum_{i=1}^{N-n}c_iM_{i,n}(y)$$

and points $0<\xi_1<\cdots<\xi_{N-n}<1$ such that

$$h(\xi_i)=(-1)^i\|h\|_\infty,\qquad i=1,\cdots,N-n.$$

Without loss of generality, we normalize h so that $\|h\|_\infty=1$. Before showing how we use h to construct a solution to (12), let us consider h and the points of equi-oscillation $(\xi_i)_1^{N-n}$ in more detail.

PROPOSITION 4.1. *Let h be as above. Then we have the following:*

(i) *h is unique.*

(ii) *$c_i(-1)^i>0$, $i=1,\cdots,N-n$.*

(iii) *The $(\xi_i)_1^{N-n}$ are uniquely defined.*

(iv) *If $n\geqq 3$, then $t_{i+1}<\xi_i<t_{i+n-1}$, $i=1,\cdots,N-n$.*

Proof. For $n=2$, h is continuous and piecewise linear with knots $t_2,\cdots,t_{N-1}$, and satisfies $h(0)=h(1)=0$. h is easily seen to exist and satisfy (i), (ii), and (iii) with $\xi_i=t_{i+1}$, $i=1,\cdots,N-2$.

Assume $n\geqq 3$. By construction h has at least $N-n-1$ sign changes. From Proposition 2.1 and the proof thereof, it follows that h has exactly $N-n-1$ sign changes, $h^{(n-1)}$ strictly changes sign at each t_i, $i=2,\cdots,N-1$, and $c_i(-1)^i>0$, $i=1,\cdots,N-n$.

For each $j\in\{1,\cdots,N\}$, $(M_{i,n})_{i=1,i\neq j}^{N-n}$ is a WT-system on $[0,1]$. Since h equi-oscillates at $N-n$ points, it follows that $-c_j^{-1}\sum_{i=1,i\neq j}^{N}c_iM_{i,n}$ is a best approximant to $M_{j,n}$ on $[0,1]$ in the uniform norm from span $\{M_{i,n}\}_{i=1,i\neq j}^{N-n}$. Furthermore, the error in the best approximation is exactly $|c_j|^{-1}$. If $\tilde{h}=\sum_{i=1}^{N-n}d_iM_{i,n}$ satisfies $\|\tilde{h}\|_\infty=1$, and $\tilde{h}(\eta_i)=(-1)^i$, $i=1,\cdots,N-n$ for some $0<\eta_1<\cdots<\eta_{N-n}<1$, then it follows as above that $(-1)^jd_j^{-1}=|d_j|^{-1}=|c_j|^{-1}=(-1)^jc_j^{-1}$. Thus $d_j=c_j$ for each j proving the

uniqueness of h. (A different proof of the uniqueness of h follows from the analysis in the proof of Theorem 4.2.)

Since

$$\sum_{i=1}^{N-n} c_i M_{i,n}(\xi_j) = (-1)^j, \qquad j = 1, \cdots, N-n,$$

and $(M_{i,n}(\xi_j))_{i,j=1}^{N-n}$ is TP, it follows that this matrix is nonsingular and thus $\xi_i \in (t_i, t_{i+n})$, $i = 1, \cdots, N-n$. However, we wish to prove more, namely, $\xi_i \in (t_{i+1}, t_{i+n-1})$, $i = 1, \cdots, N-n$. To this end we use Rolle's theorem and the fact that $h'(\xi_i) = 0$, $i = 1, \cdots, N-n$. Since $n \geqq 3$, h' is continuous and vanishes on $(0, \xi_i]$ at the points $\xi_1, \cdots, \xi_i$. Furthermore $h^{(j)}(0) = 0, j = 0, 1, \cdots, n-2$. When we apply Rolle's theorem, it follows that $h^{(n-2)}$ has at least $i+1$ distinct zeros in $[0, \xi_i]$, and $h^{(n-1)}$ has at least i sign changes in $(0, \xi_i)$ (since $h^{(n-1)}$ does not vanish identically on any subinterval). But $h^{(n-1)}$ changes sign exactly at $t_2, \cdots, t_{N-1}$. Thus $t_{i+1} < \xi_i$. Similarly we prove that $\xi_i < t_{i+n-1}$.

It remains to prove (iii) for $n \geqq 3$. Property (iii) follows if we can show that h' has no zeros in $(0, 1)$ other than $(\xi_i)_1^{N-n}$. This fact may be proven by a simple Rolle's theorem argument. Alternatively, we can argue as follows:

$$h'(y) = \sum_{i=1}^{N-n+1} d_i M_{i,n-1}(y)$$

where $\operatorname{supp} M_{i,n-1} = (t_i, t_{i+n-1})$, $i = 1, \cdots, N-n+1$. If $h'(\xi) = 0$ for some $\xi \in (0, 1) \backslash \{\xi_1, \cdots, \xi_{N-n}\}$, then by setting $\{\eta_1, \cdots, \eta_{N-n+1}\} = \{\xi_1, \cdots, \xi_{N-n}, \xi\}$ where $0 < \eta_1 < \cdots < \eta_{N-n+1} < 1$, it follows from (iv) that $t_i < \eta_i < t_{i+n-1}$, $i = 1, \cdots, N-n+1$. But this implies that $h' \equiv 0$, which is a contradiction. □

We can now construct a unique solution to (11).

THEOREM 4.2. *Let h and $(\xi_i)_1^{N-n}$ be as given above. Set*

$$\mu_t = \sum_{j=1}^{N-n} b_j \delta_{\xi_j}$$

where the $(b_j)_1^{N-n}$ are chosen so that

$$\sum_{j=1}^{N-n} b_j M_{i,n}(\xi_j) = E_i, \qquad i = 1, \cdots, N-n.$$

Then $\|\mu_t\| = \sum_{j=1}^{N-n} |b_j|$ and μ_t is the unique solution to (11).

Proof. Let μ_t be as given above. Such a μ_t exists and is unique since $(M_{i,n}(\xi_j))_{i,j=1}^{N-n}$ is nonsingular. Since $E_i(-1)^i \sigma > 0$, $i = 1, \cdots, N-n$ for some $\sigma \in \{-1, 1\}$, fixed, it follows from the total positivity of $(M_{i,n}(\xi_j))_{i,j=1}^{N-n}$ that $b_j(-1)^j \sigma > 0$, $j = 1, \cdots, N-n$. Furthermore since $c_i(-1)^i > 0$, $i = 1, \cdots, N-n$, it is easily seen that

$$\sum_{j=1}^{N-n} |b_j| = \left| \sum_{i=1}^{N-n} c_i E_i \right| = \sum_{i=1}^{N-n} |c_i E_i|.$$

For any μ, $\|\mu\| < \infty$, satisfying

$$\int_0^1 M_{i,n}(y)\, d\mu(y) = E_i, \qquad i = 1, \cdots, N-n,$$

we have

$$\left| \sum_{i=1}^{N-n} c_i E_i \right| = \left| \int_0^1 h(y)\, d\mu(y) \right| \leqq \|h\|_\infty \|\mu\|.$$

Since $\|h\|_\infty = 1$,

$$\|\mu_{\mathbf{t}}\| = \sum_{j=1}^{N-n} |b_j| = \left| \sum_{i=1}^{N-n} c_i E_i \right| \leqq \|\mu\|,$$

which implies that $\mu_{\mathbf{t}}$ is a solution to (11). Assume $\|\mu_{\mathbf{t}}\| = \|\mu\|$. Then necessarily

$$\left| \int_0^1 h(y)\, d\mu(y) \right| = \|h\|_\infty \|\mu\|.$$

By Proposition 4.1(iii), h attains its norm only at the values $(\xi_j)_1^{N-n}$. Thus μ has support only on the $(\xi_j)_1^{N-n}$. Since $(M_{i,n}(\xi_j))_{i,j=1}^{N-n}$ is nonsingular, it follows that $\mu \equiv \mu_{\mathbf{t}}$. □

We now turn our attention to the problem as stated in (4), namely,

$$\min_{\mathbf{t} \in \Xi_N} \min \{\|\mu\| : f \in V^{(n)}(\mathbf{t}; \mathbf{e})\}. \tag{13}$$

From Theorem 4.2 we know that any solution must be of the specific form given therein. We will prove that any solution must also oscillate strictly between the $(e_i)_1^N$ on some $(t_i)_1^N$. For $n = 2$, this result is simple, and yet disappointing. For any $\mathbf{t} \in \Xi_N$, $t_1 = 0 < t_2 < \cdots < t_N = 1$, construct $\mu_{\mathbf{t}}$ and the associated $f(x) = a_0 + a_1 x + \sum_{j=1}^{N-2} b_j (x - \xi_j)_+^1$. As noted earlier $\xi_j = t_{j+1}, j = 1, \cdots, N-2$. Since $(e_i - e_{i-1}) \times (e_{i+1} - e_i) < 0$, $i = 2, \cdots, N-1$, it is easily seen that for any choice of $\mathbf{t} \in \Xi_N$, f oscillates strictly between the $(e_i)_1^N$ on $(t_i)_1^N$. Thus our result clearly holds, but obviously this condition is not sufficient in (13). However, it is possible, by calculation, to verify that the solution to (13) is unique.

We now turn our attention to the case $n \geqq 3$. Here it is unclear as to whether the following necessary conditions are also sufficient for a solution to (13), and also as to whether uniqueness holds.

THEOREM 4.3. *Let $n \geqq 3$ and let f^* be any solution to* (13). *There exists a $\mathbf{t}^* = (t_1^*, \cdots, t_N^*)$, $0 = t_1^* < \cdots < t_N^* = 1$ such that $f^* \in V^{(n)}(\mathbf{t}^*; \mathbf{e})$. Furthermore,*

(a) $$f^*(x) = \sum_{i=0}^{n-1} a_i x^i + \sum_{j=1}^{N-n} b_j (x - \xi_j)_+^{n-1}$$

where the $(b_j)_1^{N-n}$ and $(\xi_j)_1^{N-n}$ are as in Theorem 4.2 *with respect to $(t_i^*)_1^N$.*

(b) *f^* oscillates strictly between the $(e_i)_1^N$ on $(t_i^*)_1^N$.*

Proof. Let f^* be a solution to (13). There then exists a $\mathbf{t}^*$ as above for which $f^* \in V^{(n)}(\mathbf{t}^*; \mathbf{e})$. Since f^* must also solve (11) for $\mathbf{t}^*$, it follows that f^* necessarily satisfies (a). It remains to prove (b).

We first prove that $f^{*\prime}(t_i^*) = 0$, $i = 2, \cdots, N-1$. Assume to the contrary that $f^{*\prime}(t_k^*) \neq 0$ for some $k \in \{2, \cdots, N-1\}$. Without loss of generality we will assume that $(e_k - e_{k-1}) > 0$. Thus there exists an s_k as near as we wish to t_k^* for which $f(s_k) > e_k$. Recall from Proposition 4.1 that $t_{i+1}^* < \xi_i < t_{i+n-1}^*$, $i = 1, \cdots, N-n$. Taking s_k close to t_k^*, this implies the existence of a unique

$$g(x) = \sum_{i=0}^{n-1} c_i x^i + \sum_{j=1}^{N-n} d_j (x - \xi_j)_+^{n-1}$$

that satisfies $g(t_i^*) = e_i$, $i = 1, \cdots, N$, $i \neq k$, and $g(s_k) = e_k$. Furthermore $\operatorname{sgn} d_j = \operatorname{sgn} b_j = (-1)^j \sigma$, $j = 1, \cdots, N-n$, where $\sigma \in \{-1, 1\}$, fixed. g is "admissible" in (13) with $\|\mu\| = \sum_{j=1}^{N-n} |d_j|$. We will prove that $\sum_{j=1}^{N-n} |d_j| < \sum_{j=1}^{N-n} |b_j|$, contradicting the minimality of f^*. To this end, note that

$$(f^* - g)(x) = \sum_{i=0}^{n-1} (a_i - c_i) x^i + \sum_{j=1}^{N-n} (b_j - d_j)(x - \xi_j)_+^{n-1}$$

satisfies $(f^*-g)(t_i^*)=0$, $i=1,\cdots,N$, $i\neq k$, and $(f^*-g)(s_k)>0$. The conditions $t_{i+1}^*<\xi_i<t_{i+n-1}^*$, $i=1,\cdots,N-n$, easily imply that f^*-g vanishes only at $(t_i^*)_{i=1,i\neq j}^N$. Thus, in particular, $(f^*-g)(t_k^*)>0$.

Set $(f^*-g)[t_i^*,\cdots,t_{i+n}^*]=F_i$, $i=1,\cdots,N-n$. Then

$$F_i=\sum_{j=1}^{N-n}(b_j-d_j)M_{i,n}(\xi_j),\qquad i=1,\cdots,N-n$$

and $F_i(-1)^i\sigma\geqq 0$, $i=1,\cdots,N-n$. Furthermore the $(F_i)_1^{N-n}$ are not all zero. From the total positivity of $(M_{i,n}(\xi_j))_{i,j=1}^{N-n}$ it follows that $(b_j-d_j)(-1)^j\sigma\geqq 0, j=1,\cdots,N-n$, and not all the $(b_j-d_j)_1^{N-n}$ are zero (in fact all are nonzero). Thus

$$\sum_{j=1}^{N-n}|b_j|=\sum_{j=1}^{N-n}b_j(-1)^j\sigma>\sum_{j=1}^{N-n}d_j(-1)^j\sigma=\sum_{j=1}^{N-n}|d_j|.$$

Therefore $f^{*\prime}(t_i^*)=0$, $i=2,\cdots,N-1$.

It remains to prove that f^* is strictly monotone on $[t_i^*,t_{i+1}^*]$ for each $i=1,\cdots,N-1$. Using the fact that $t_{i+1}^*<\xi_i<t_{i+n-1}^*$, $i=1,\cdots,N-n$, it follows that $f^{*\prime}(x)$ has no zeros in $[0,1]$ other than $t_2^*,\cdots,t_{N-1}^*$. Thus f^* oscillates strictly between the $(e_i)_1^N$ on $(t_i^*)_1^N$. □

Acknowledgments. The author thanks Professors B. D. Bojanov and K. Scherer for their help.

REFERENCES

[1] B. D. Bojanov, *Perfect splines of least uniform deviation*, Anal. Math., 6 (1980), pp. 185-197.
[2] C. de Boor, *On "Best" interpolation*, J. Approx. Theory, 16 (1976), pp. 28-42.
[3] ———, *Splines as linear combinations of* B-*splines: a survey*, in Approximation Theory II, G. G. Lorentz, C. K. Chui, and L. L. Schumaker, eds., Academic Press, New York, 1976, pp. 1-47.
[4] C. Davis, *Extrema of a polynomial. Advanced problem.* Amer. Math. Monthly, 64 (1957), pp. 679-680.
[5] S. D. Fisher and J. W. Jerome, *Spline solutions to L^1 extremal problems in one and several variables*, J. Approx. Theory, 13 (1975), pp. 73-83.
[6] ———, *Minimum Norm Extremals in Function Spaces*, Lecture Notes in Mathematics 479, Springer-Verlag, Berlin, 1975.
[7] S. Karlin, *Total Positivity, Vol.* I. Stanford University Press, Stanford, CA, 1968.
[8] ———, *Interpolation properties of generalized perfect splines and the solutions of certain extremal problems*, Trans. Amer. Math. Soc., 106 (1975), pp. 25-66.
[9] S. P. Marin, *An approach to data parametrization in parametric cubic spline interpolation problems*, J. Approx. Theory, 41 (1984), pp. 64-86.
[10] J. Mycielski and S. Paszkowski, *A generalization of Chebyshev polynomials*, Bull. Acad. Polon. Sci. Sér. Sci. Math., 8 (1960), pp. 433-438.
[11] K. Scherer and P. W. Smith, *Remarks on best interpolation by curves*, preprint.
[12] H.-J. Töpfer, *Models for smooth curve fitting*, in Numerical Methods of Approximation Theory, Vol. 6, Birkhäuser-Verlag, Basel, 1982, pp. 209-224.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

016

EXTENSION OF SZEGÖ'S THEOREM ON THE SECTIONS OF UNIVALENT FUNCTIONS*

STEPHAN RUSCHEWEYH†

Abstract. In this paper a far-reaching extension of Szegö's theorem on the univalence of partial sums of the power series expansion of univalent functions in the class $\mathcal{S}$ is given. In particular, it is shown that the property "univalent" can be replaced by the stronger one "starlike univalent" and that the conclusion is not only true for $\mathcal{S}$ but also for the closed convex hull of $\mathcal{S}$. The paper concludes with the discussion of a new conjecture on $\mathcal{S}$, stronger than the former "Bieberbach conjecture."

Key words. univalent functions, partial sums, Hadamard product

AMS(MOS) subject classification. 30C45

1. Introduction and statement of the results. Let $\mathcal{A}$ denote the set of analytic functions f in the unit disk $\mathbb{D}$ satisfying $f(0)=0, f'(0)=1$, and let $\mathcal{S}$ be the set of univalent functions in $\mathcal{A}$. For

$$f(z)=\sum_{k=1}^{\infty} a_k z^k \in \mathcal{A},$$

let

$$f_n(z)=\sum_{k=1}^{n} a_k z^k \qquad (n \in \mathbb{N}).$$

In 1928, G. Szegö [5] proved the following nice result.

THEOREM A. *For $f \in \mathcal{S}$ all sections f_n are univalent in $\mathbb{D}_{1/4}$.*

Here $\mathbb{D}_R$ denotes the set $\{z: |z|<R\}$. A function $f \in \mathcal{A}$ is said to be convex (starlike) in $\mathbb{D}_R$ if it is univalent in $\mathbb{D}_R$ with $f(\mathbb{D}_R)$ convex (starlike with respect to the origin). By $\mathcal{C}$, $\mathcal{S}^*$ we denote the subclasses of functions in $\mathcal{S}$ which are convex or starlike in $\mathbb{D}$, respectively. Szegö also proved the following theorem.

THEOREM B. *If $f \in \mathcal{C}$ ($\mathcal{S}^*$), then all sections f_n are convex (starlike) in $\mathbb{D}_{1/4}$.*

In the present paper we show that both Theorems A and B are very special cases of a general convolution theorem. We recall that for

$$f(z)=\sum_{k=0}^{\infty} a_k z^k, \qquad g(z)=\sum_{k=0}^{\infty} b_k z^k$$

the Hadamard product (or convolution) is defined by

$$(f * g)=\sum_{k=0}^{\infty} a_k b_k z^k$$

and that $\mathcal{A}$ is closed under convolution, i.e., $f, g \in \mathcal{A}$ implies $f * g \in \mathcal{A}$.

Let

$$\mathcal{F} := \operatorname{clco}\left\{\sum_{k=1}^{n} x^{k-1} z^k : |x| \leqq 1\right\}$$

* Received by the editors March 5, 1986; accepted for publication (in revised form) January 11, 1987.

† Math. Institut d. Universität, D-8700 Würzburg, Federal Republic of Germany. Current address, Depto. Matemáticas, Universidad Técnica F.S.M. Casilla 110-V, Valparaiso, Chile.

where clco stands for the closed convex hull (in this case with respect to the linear space of analytic functions in $\mathbb{D}$ with the topology of compact convergence in $\mathbb{D}$). $\mathcal{F}$ contains two interesting subsets:

$$\mathcal{R} := \{f \in \mathcal{A}: \operatorname{Re}(f(z)/z) > \tfrac{1}{2}, z \in \mathbb{D}\} \subset \mathcal{F}, \tag{1.1}$$

$$\mathcal{Q} := \left\{f(z) = \sum_{k=1}^{\infty} a_k z^k \in \mathcal{A}: 0 \leqq a_{k+1} \leqq a_k, k \in \mathbb{N}\right\} \subset \mathcal{F}. \tag{1.2}$$

In particular, since $\mathcal{C} \subset \mathcal{R}$ (compare Duren [2, pp. 72-73]), we have

$$\mathcal{C} \subset \mathcal{F}. \tag{1.3}$$

We will show that $\mathcal{F}$ is closed under convolutions:

$$f, g \in \mathcal{F} \Rightarrow f * g \in \mathcal{F} \tag{1.4}$$

and, in particular,

$$f \in \mathcal{F} \Leftrightarrow f_n \in \mathcal{F} \quad \text{for } n \in \mathbb{N}. \tag{1.5}$$

Our main result is as follows.

THEOREM 1. *Let $f \in \operatorname{clco} \mathcal{S}$ and $g \in \mathcal{F}$. Then $f * g$ is starlike in $\mathbb{D}_{1/4}$. The constant $\frac{1}{4}$ is best possible.*

This obviously contains both Theorems A and B using the special choices $g = s_n$ where

$$s_n(z) = \sum_{k=1}^{n} z^k \qquad (n \in \mathbb{N}).$$

But, in fact, Theorem A is considerably improved by Theorem 1; the result is true not only for $f \in \mathcal{S}$ but for all $f \in \operatorname{clco} \mathcal{S}$, a much larger set, which includes, for instance, all normalized typically real functions in $\mathbb{D}$; and the conclusion "univalent" is replaced by a stronger one, "starlike".[1] Also Theorem B is extended by Theorem 1, as can be seen from Corollary 1 and the above-mentioned inclusion $\mathcal{C} \subset \mathcal{R} \subset \mathcal{F}$.

COROLLARY 1. *If $f \in \mathcal{F}$ then f is convex in $\mathbb{D}_{1/4}$. In particular, f_n is convex in $\mathbb{D}_{1/4}$ for $n \in \mathbb{N}$. The constant $\frac{1}{4}$ is best possible.*

Note that this also gives the sharp "radius of convexity" within the class $\mathcal{Q} \subset \mathcal{F}$: the function $s_2 \in \mathcal{Q}$ is convex in $\mathbb{D}_{1/4}$ but in no larger disk.

Theorem 1 is related to a general convolution conjecture we are proposing. Let

$$\mathcal{D} = \{g \in \mathcal{A}: |g''(z)| \leqq \operatorname{Re} g'(z), z \in \mathbb{D}\}.$$

Conjecture. Let $f \in \operatorname{clco} \mathcal{S}$, $g \in \mathcal{D}$. Then $f * g \in \mathcal{S}^*$.

We can prove that $h \in \mathcal{F}$ implies $g(z) := 4h(z/4) \in \mathcal{D}$ and this shows that Theorem 1 is a partial verification of the conjecture.

Regarding $\mathcal{D}$ we prove the following theorem.

THEOREM 2. (i) *$\mathcal{D}$ is a compact convex subset of $\mathcal{C}$, closed under convolutions.*

(ii) *Let $g(z) = \sum_{k=1}^{\infty} a_k z^k \in \mathcal{A}$ with*

$$\sum_{k=2}^{\infty} k^2 |a_k| \leqq 1. \tag{1.6}$$

Then $g \in \mathcal{D}$ and the conjecture is true for g.

[1] After completing this manuscript I became aware of the paper entitled, "On a theorem of Szegö," by K. Hu and Y. F. Pan, *J. Math. Res. Exposition*, 4 (1984), pp. 41-44. The authors also show that "univalent" can be replaced by "starlike univalent" in the conclusion of Theorem A. Their method, however, does not extend to clco $\mathcal{S}$.

We also note that the conjecture is stronger than the (former) Bieberbach conjecture (de Branges' theorem). In fact, assume

$$f(z)=\sum_{k=1}^{\infty} a_k z^k \in \mathcal{S}$$

and that the conjecture holds. Since, by Theorem 2,

$$g_n(z)=z+z^n/n^2 \in \mathcal{D},$$

we see that $(f * g_n)(z)=z+a_n z^n/n^2 \in \mathcal{S}^*$, which gives $|a_n| \leqq n$, $n=2,3,\cdots$.

2. Proofs. We first verify the claims (1.1)-(1.5). Since $\mathcal{F}$ is closed by definition, it contains the functions

$$f_x(z) := \frac{z}{1-xz}, \qquad |x|=1,$$

and hence also clco $\{f_x: |x|=1\}$, which is known to equal $\mathcal{R}$. This proves (1.1).

For the proof of (1.2) we note that the set of polynomials in $\mathcal{Q}$ is dense in $\mathcal{Q}$, and since $\mathcal{F}$ is closed it suffices to show that every such polynomial is in $\mathcal{F}$. But these polynomials are finite convex combinations of the $s_n \in \mathcal{F}$, $n \in \mathbb{N}$, and $\mathcal{F}$ is convex.

Let $1 \leqq n \leqq m$ and

$$f(z)=\sum_{k=1}^{n} x^{k-1} z^k, \qquad g(z)=\sum_{k=1}^{m} y^{k-1} z^k$$

for certain $x, y \in \bar{\mathbb{D}}$. Then

$$(f * g)(z)=\sum_{k=1}^{n} (xy)^{k-1} z^k \in \mathcal{F}.$$

Since $\mathcal{F}$ is obviously compact we can apply [3, Thm. 1.17] (with a slight modification concerning the normalization) to deduce (1.4). Then from (1.4) with $g=s_n$ and $f \in \mathcal{F}$ we have $f * s_n = f_n \in \mathcal{F}$, $n \in \mathbb{N}$, which is one direction of (1.5). The other direction follows from the compactness of $\mathcal{F}$.

For the proof of Theorem 1 we require a fairly large number of results on $\mathcal{S}$ which we state in the following three theorems. Except for (2.5) (de Branges' theorem [1]), these results can be found in the standard textbooks on univalent functions (see, for example, Duren [2]).

THEOREM C. *For $f \in \mathcal{S}$ and $z \in \mathbb{D}$ we have*

$$\left|\log \frac{f(z)}{z}+\log (1-|z|^2)\right| \leqq \log \frac{1+|z|}{1-|z|}, \tag{2.1}$$

$$\left|\log \frac{zf'(z)}{f(z)}\right| \leqq \log \frac{1+|z|}{1-|z|}, \tag{2.2}$$

$$|\arg f'(z)| \leqq 4 \arcsin |z| \qquad (|z|<1/\sqrt{2}), \tag{2.3}$$

$$|f'(z)| \geqq \frac{1-|z|}{(1+|z|)^3}. \tag{2.4}$$

THEOREM D. *Let $f(z)=\sum_{k=1}^{\infty} a_k z^k \in \mathcal{S}$. Then*

$$|a_n| \leqq n \qquad (n \in \mathbb{N}), \tag{2.5}$$

$$|a_3-a_2^2| \leqq 1. \tag{2.6}$$

THEOREM E. *Let*

$$V_3 = \{(a_2, a_3) \in \mathbb{C}^2 : \exists f(z) = z + a_2 z^2 + a_3 z^3 + \cdots \in \mathcal{S}\}.$$

Then the set of points (a_2, a_3) *with*

$$a_2 = -2\int_0^1 x(t)\,dt, \qquad a_3 = a_2^2 - 2\int_0^1 t x^2(t)\,dt,$$

where $x(t)$ *is continuous with* $|x(t)| = 1$, *is dense in* V_3.

We now give a series of lemmas based on these theorems. The results derived in these lemmas are not sharp but sufficient for our purposes. Let

$$\beta = \log \tfrac{5}{3} = 0.5108\cdots.$$

We note that $e^{\beta z}$ is a (not normalized) convex univalent function in $\bar{\mathbb{D}}$.

LEMMA 1. *We have*

$$\left|\frac{1 - e^{\beta z}}{1 + e^{\beta z}}\right| \leqq \sigma \qquad (z \in \bar{\mathbb{D}}), \tag{2.7}$$

where $\sigma = \sqrt{2}/3\,(1 + \cos\beta)^{-1/2} = 0.3445\cdots$.

Proof. Let $f(z) = (1 - e^{\beta z})/(1 + e^{\beta z})$. Then f is analytic in $\bar{\mathbb{D}}$ since $\operatorname{Re} e^{\beta z} > 0$, $z \in \bar{\mathbb{D}}$. It therefore suffices to prove (2.7) on $|z| = 1$. Also, since

$$|F(z)| = |F(-z)| = |F(\bar{z})|,$$

we can restrict our attention to $z = e^{i\theta}$, $0 \leqq \theta \leqq \pi/2$. But in that range we have

$$|e^{\beta z}| = e^{\beta \cos\theta} \geqq 1,$$

and, since $e^{\beta z}$ is convex univalent in $\bar{\mathbb{D}}$, we see that in the same range $\operatorname{Re} e^{\beta z} \geqq \operatorname{Re} e^{i\beta}$. Simple geometry now shows that

$$|1 + e^{\beta z}| \geqq |1 + e^{i\beta}| = \sqrt{2(1 + \cos\beta)},$$

which together with the trivial estimate

$$|1 - e^{\beta z}| \leqq e^{\beta} - 1 = \tfrac{2}{3}$$

gives (2.7).

LEMMA 2. *Let* $f \in \mathbb{D}_{1/4}$. *Then*

$$\left|f'(z) - \frac{f(z)}{z}\right| \leqq \alpha \operatorname{Re}\left(f'(z) + \frac{f(z)}{z}\right), \tag{2.8}$$

where $\alpha < 0.53$.

Proof. From (2.2) we conclude that in $\mathbb{D}_{1/4}$

$$\frac{zf'(z)}{f(z)} = e^{\rho(z)\beta} \tag{2.9}$$

where $|\rho(z)| \leqq 1$. Hence, by Lemma 1,

$$\left|f'(z) - \frac{f(z)}{z}\right| \leqq \sigma \left|f'(z) + \frac{f(z)}{z}\right| \qquad (z \in \mathbb{D}_{1/4}). \tag{2.10}$$

Now we estimate the entity

$$\frac{|f(z) + f(z)/z|}{\operatorname{Re}\,(f'(z) + f(z)/z)} = \frac{1}{\cos\,(\arg\,(f'(z) + f(z)/z))}. \tag{2.11}$$

From (2.1) we obtain

$$|\arg f(z)/z| \leqq \beta \qquad (z \in \mathbb{D}_{1/4}). \tag{2.12}$$

It is the consequence of Lemma 1 and (2.9) that

$$1+\frac{zf'(z)}{f(z)} = 1+\frac{1+\sigma^2}{1-\sigma^2}+\varepsilon(z)\frac{2\sigma}{1-\sigma^2}$$

with $|\varepsilon(z)| \leqq 1$, and therefore

$$\left|\arg\left(1+\frac{zf'(z)}{f(z)}\right)\right| = |\arg(1+\varepsilon(z)\sigma)| \leqq \arcsin\sigma.$$

This together with (2.12) shows that the function (2.11) is bounded by

$$1/\cos(\beta+\arcsin\sigma) < \frac{0.53}{\sigma}, \tag{2.13}$$

and a combination of (2.10), (2.11), and (2.13) yields (2.8).

LEMMA 3. *For $f \in \mathcal{S}$ and $z \in \mathbb{D}_{1/4}$ we have*

$$\operatorname{Re}\left(f'(z)+\frac{f(z)}{z}\right) \geqq \frac{211}{250}. \tag{2.14}$$

Proof. From (2.1), with $|z| = \frac{1}{4}$, we obtain

$$\left|\log\left(\frac{15}{16}\frac{f(z)}{z}\right)\right| \leqq \beta$$

which extends, by the maximum principle, to $|z| \leqq \frac{1}{4}$. Hence there exists an analytic function ρ in $\mathbb{D}_{1/4}$ with

$$\frac{f(z)}{z} = \frac{16}{15}e^{\rho(z)\beta} \qquad (|\rho(z)| \leqq 1)$$

and, again using the convexity of $e^{\beta z}$ in $\mathbb{D}$, we get

$$\operatorname{Re}\frac{f(z)}{z} \geqq \frac{16}{15}e^{-\beta} = \frac{16}{25} \qquad (z \in \mathbb{D}_{1/4}). \tag{2.15}$$

On the other hand,

$$\operatorname{Re} f'(z) = |f'(z)|\cos(\arg f'(z))$$

and a combination of (2.3), (2.4) yields

$$\operatorname{Re} f'(z) \geqq \frac{1-\frac{1}{4}}{(1+\frac{1}{4})^3}\cos\left(4\arcsin\frac{1}{4}\right) = \frac{51}{250} \qquad (z \in \mathbb{D}_{1/4}). \tag{2.16}$$

Formulae (2.15) and (2.16) give (2.14).

LEMMA 4. *Let $(a_2, a_3) \in V_3$. Then*

$$|2a_2+a_3| \leqq 16+6\operatorname{Re} a_2+2\operatorname{Re} a_3. \tag{2.17}$$

Proof. We show first that

$$\operatorname{Re}(3a_2+a_3) \geqq -\frac{21}{4}. \tag{2.18}$$

Using the representations of Theorem E we obtain

$$\begin{aligned}
\operatorname{Re}(3a_2+a_3) &= \operatorname{Re}\left\{-6\int_0^1 x(t)\,dt+4\left(\int_0^1 x(t)\,dt\right)^2-2\int_0^1 tx^2(t)\,dt\right\}\\
&= -\frac{9}{4}+4\operatorname{Re}\left(\int_0^1\left(x(t)-\frac{3}{4}\right)dt\right)^2-2\operatorname{Re}\int_0^1 tx^2(t)\,dt\\
&= -\frac{5}{4}+4\left(\int_0^1 \operatorname{Re}\left(x(t)-\frac{3}{4}\right)dt\right)^2-4\left(\int_0^1 \operatorname{Im} x(t)\,dt\right)^2\\
&\quad -4\int_0^1 t(\operatorname{Re} x(t))^2\,dt\\
&\geqq -\frac{5}{4}+4\left(\int_0^1 \operatorname{Re}\left(x(t)-\frac{3}{4}\right)dt\right)^2-4\int_0^1 (\operatorname{Im} x(t))^2\,dt\\
&\quad -4\int_0^1 t(\operatorname{Re} x(t))^2\,dt\\
&= -\frac{21}{4}+4\left(\int_0^1 \operatorname{Re}\left(x(t)-\frac{3}{4}\right)dt\right)^2+4\int_0^1 (1-t)(\operatorname{Re} x(t))^2\,dt\\
&\geqq -\frac{21}{4}.
\end{aligned}$$

Here we made use of the Cauchy-Schwarz inequality and the relations

$$\operatorname{Re} A^2=(\operatorname{Re} A)^2-(\operatorname{Im} A)^2=-|A|^2+2(\operatorname{Re} A)^2 \qquad (A\in\mathbb{C}).$$

When we use the bounds $|a_2|\leqq 2, |a_3|\leqq 3$ it follows immediately that (2.17) holds if

$$\operatorname{Re} a_2\geqq -\frac{1}{2}.$$

Also, by (2.18) we deduce that (2.17) holds if

$$|2a_2+a_3|\leqq 16-\frac{21}{2}=\frac{11}{2} \tag{2.19}$$

which is, in particular, the case if

$$|a_2|\leqq\frac{5}{4}.$$

Now let $\operatorname{Re} a_2<-\frac{1}{2}, |a_2|>\frac{5}{4}$ and write $a_3=a_2^2+\rho$ where $|\rho|\leqq 1$ by (2.6). Then we have

$$\begin{aligned}
|2a_2+a_3|^2 &= 4|a_2|^2+|a_3|^2+4\operatorname{Re} a_2\overline{a_3}\\
&\leqq 25+4\operatorname{Re} a_2(\overline{a_2}^2+\bar{\rho})\\
&\leqq 25-4\left(\frac{5}{4}\right)^2\cdot\frac{1}{2}+8\\
&= \frac{239}{8}<\left(\frac{11}{2}\right)^2.
\end{aligned}$$

This shows that also in this case (2.19), and hence (2.17) holds.

LEMMA 5. *Let $f \in \mathcal{S}$. Then for $n \in \mathbb{N}$ and $z \in \mathbb{D}_{1/4}$ we have*

$$\left| f_n'(z) - \frac{f_n(z)}{z} \right| \leqq \operatorname{Re}\left(f_n'(z) + \frac{f_n(z)}{z} \right). \tag{2.20}$$

Proof. We write $f_n = f - p_n$ such that by Lemma 2 for $z \in \mathbb{D}_{1/4}$

$$\begin{aligned} \left| f_n'(z) - \frac{f_n(z)}{z} \right| &\leqq \left| f'(z) - \frac{f(z)}{z} \right| + \left| p_n'(z) - \frac{p_n(z)}{z} \right| \\ &\leqq \alpha \operatorname{Re}\left(f'(z) + \frac{f(z)}{z} \right) + \left| p_n'(z) - \frac{p_n(z)}{z} \right|. \end{aligned}$$

In order to estimate the right-hand side using

$$\operatorname{Re}\left(f_n'(z) + \frac{f_n(z)}{z} \right) = \operatorname{Re}\left(f'(z) + \frac{f(z)}{z} \right) - \operatorname{Re}\left(p_n'(z) + \frac{p_n(z)}{z} \right)$$

it suffices to show that

$$\left| p_n'(z) - \frac{p_n(z)}{z} \right| + \left| p_n'(z) + \frac{f_n(z)}{z} \right| \leqq (1-\alpha) \operatorname{Re}\left(f'(z) + \frac{f(z)}{z} \right).$$

From (2.5) we obtain

$$\left| p_n'(z) \pm \frac{p_n(z)}{z} \right| \leqq \sum_{k=n+1}^{\infty} \frac{k^2 \pm k}{4^{k-1}} \qquad (z \in \mathbb{D}_{1/4})$$

and, using Lemma 3, we are left with the inequality

$$\sum_{k=n+1}^{\infty} \frac{k^2}{4^{k-1}} \leqq \frac{0.47}{2} \cdot \frac{211}{250} = 0.198 \cdots.$$

A simple calculation shows that this is true for $n \geqq 4$. Formula (2.20) is immediately verified for $n = 2$. For $n = 3$ we have to show that

$$|a_2 z + 2a_3 z^2| \leqq 2 + \operatorname{Re}(3a_2 z + 4a_3 z^2) \qquad (z \in \mathbb{D}_{1/4}). \tag{2.21}$$

Since $(a_2, a_3) \in V_3$ implies $(a_2 x, a_3 x^2) \in V_3$ for $x \in \bar{\mathbb{D}}$ we see that (2.21) is equivalent to

$$\left| \frac{a_2}{4} + \frac{a_3}{8} \right| \leqq 2 + \operatorname{Re}\left(\frac{3}{4} a_2 + \frac{1}{4} a_3 \right)$$

for $(a_2, a_3) \in V_3$. This, however, is the content of Lemma 4.

Proof of Theorem 1. Let $\mathcal{H}_R$ denote the class of functions $h \in \mathcal{A}$ with

$$\left| h'(z) - \frac{h(z)}{z} \right| \leqq \operatorname{Re}\left(h'(z) + \frac{h(z)}{z} \right) \qquad (z \in \mathbb{D}_R). \tag{2.22}$$

Then (2.22) implies

$$\left| \left(\frac{zh'(z)}{h(z)} - 1 \right) \Big/ \left(\frac{zh'(z)}{h(z)} + 1 \right) \right| \leqq 1 \qquad (z \in \mathbb{D}_R),$$

and hence $\operatorname{Re}(zh'(z)/h(z)) > 0$ in $\mathbb{D}_R$. Thus (2.22) implies the starlikeness of $h \in \mathcal{H}_R$ in $\mathbb{D}_R$. We also note that $\mathcal{H}_R$ is convex and compact (this class was introduced by Singh [4]). In Lemma 5 we have seen that for $f \in \mathcal{S}$ and

$$g(z) = \sum_{k=1}^{n} x^{k-1} z^k \qquad (|x| \leqq 1, n \in \mathbb{N})$$

we have

$$f * g \in \mathcal{H}_{1/4}. \tag{2.23}$$

By the convexity and compactness of $\mathscr{H}_{1/4}$ and the linearity of the convolution we deduce that (2.23) holds also for $f \in \operatorname{clco} \mathscr{S}$, $g \in \mathscr{F}$, which is an even stronger result than that which we have claimed in Theorem 1. That the constant $\frac{1}{4}$ is best possible is readily seen from the example

$$f(z) = \frac{z}{(1-z)^2} \in \mathscr{S}, \qquad g(z) = z + z^2 \in \mathscr{F}.$$

Then $(f * g)(z) = z + 2z^2$ is starlike in $\mathbb{D}_{1/4}$ but in no larger disk.

It follows from Theorem 1 that for $f \in \mathscr{F}$ the function $z/(1-z)^2 * f = zf'(z)$ is starlike in $\mathbb{D}_{1/4}$. But this is equivalent to the assertion of Corollary 1.

Proof of Theorem 2. The functions $g \in \mathbb{D}$ fulfill, in particular, $|zg''(z)| \leqq |g'(z)|$ and therefore $\operatorname{Re}(zg''(z)/g'(z)) + 1 > 0$ in $\mathbb{D}$. Hence $\mathscr{D} \subset \mathscr{C}$, a compact set. That $\mathscr{D}$ is closed and convex is obvious. Now let $f, g \in \mathscr{D}$. Then $g \in \mathscr{C} \subset \mathscr{R}$ and therefore

$$g(z) = \int_{|x|=1} \frac{z}{1-xz}\, d\mu$$

for a certain probability measure μ on $|x| = 1$. Thus

$$(f * g)(z) = \int_{|x|=1} \frac{1}{x} f(xz)\, d\mu \in \operatorname{clco} \mathscr{D} = \mathscr{D},$$

which shows that $\mathscr{D}$ is closed under convolutions. A simple calculation shows that a function

$$h(z) = z + \alpha z^n \tag{2.24}$$

belongs to $\mathscr{D}$ if and only if $|\alpha| \leqq 1/n^2$. The functions (1.6) are in the closed convex hull of the functions (2.24) with $|\alpha| \leqq 1/n^2$ and thus in $\mathscr{D}$. Let g satisfy (1.6) and

$$f(z) = \sum_{k=1}^{\infty} b_k z^k \in \operatorname{clco} \mathscr{S}.$$

Then

$$f * g = \sum_{k=1}^{\infty} a_k b_k z^k$$

satisfies (using (2.5))

$$\sum_{k=2}^{\infty} k|a_k b_k| \leqq \sum_{k=2}^{\infty} k^2 |a_k| \leqq 1,$$

which is the well-known sufficient condition for $f * g \in \mathscr{S}^*$.

REFERENCES

[1] L. DE BRANGES, *A proof of the Bieberbach conjecture*, Acta Math., 154 (1985), pp. 137-152.

[2] P. DUREN, *Univalent Functions*, Springer-Verlag, Berlin, New York, 1983.

[3] S. RUSCHEWEYH, *Convolutions in Geometric Function Theory*, Sém. Math. Sup. 83, University of Montréal, Montréal, Québec, Canada 1982.

[4] V. SINGH, *Convolution operators on some classes of functions analytic in the unit disk*, unpublished.

[5] G. SZEGÖ, *Zur Theorie der schlichten Abbildungen*, Math. Ann., 100 (1928), pp. 188-211.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

017

ON THE ZEROS OF DERIVATIVES OF BESSEL FUNCTIONS*

LEE LORCH† AND PETER SZEGO‡

Abstract. Bessel functions of the first and second kind are denoted as usual by $J_\nu(x)$, $Y_\nu(x)$, the general cylinder function $AJ_\nu(x)+BY_\nu(x)$, where A, B are independent of x and ν, by $C_\nu(x)$, their respective positive zeros and those of their derivatives by $j_{\nu k}$, $y_{\nu k}$, $c_{\nu k}$, $j'_{\nu k}$, $y'_{\nu k}$, $c'_{\nu k}$, etc. It is shown here that for $-1<\nu<0$, (i) $j'_{\nu k}$ increases in ν, $k=1,2,\cdots$, (ii) $j'_{\nu 1}>j'_{11}=1.84\cdots$, and (iii) $(-1)^k J''_\nu(j'_{\nu k})>0$. It is also established that $c'_{\nu k}-c_{\nu l}$ increases in $\nu>0$, provided $c'_{\nu k}>c_{\nu l}>\nu>0$, where the ranks k, l may or may not be equal but are kept fixed as ν varies. Further, (iii') $(-1)^k J''_\nu(j'_{\nu k})<0$ for $0<\nu\leqq 1$, $k=1,2,\cdots$; $J''_0(j'_{01})=0$, $(-1)^k J''_0(j'_{0k})<0$, $k=2,3,\cdots$.

Key words. Bessel functions, zeros

AMS(MOS) subject classification. 33A

1. Introduction. The Bessel function of the first kind and order ν is defined to be [2, p. 4 (2)]

$$J_\nu(x)=2^{-\nu}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+\nu}}{4^m m!\Gamma(m+\nu+1)}. \tag{1}$$

It satisfies the differential equations [2, p. 4 (1); p. 13 (67)]

$$x^2y''+xy'+(x^2-\nu^2)y=0, \tag{2}$$

and

$$x^2(x^2-\nu^2)y'''+x(x^2-3\nu^2)y''+[(x^2-\nu^2)^2-(x^2+\nu^2)]y'=0. \tag{3}$$

Here we shall consider only real values of x, y, and ν, and concern ourselves chiefly with the interval $-1<\nu<0$. The kth positive zero of $J_\nu(x)$ is denoted by $j_{\nu k}$, of $J'_\nu(x)$ by $j'_{\nu k}$, except that $j'_{01}=0$. For the standard second solution of (2), $Y_\nu(x)$, and its derivative, positive zeros are denoted by $y_{\nu k}$, $y'_{\nu k}$, respectively.

General solutions $C_\nu(x)$ of (2) will also be considered here, in the form

$$C_\nu(x)=AJ_\nu(x)+BY_\nu(x),$$

where the constants A, B are independent of both x and ν. The positive zeros of $C_\nu(x)$ and $C'_\nu(x)$ are denoted by $c_{\nu k}$, $c'_{\nu k}$, respectively, and generically by c_ν, c'_ν or simply by c, c'.

It is well known, for each fixed $k=1,2,\cdots$, that $j_{\nu k}$ increases with ν, provided $\nu>-1$ [7, p. 508] and that $j'_{\nu k}$ increases with ν provided that $\nu>0$ [5, p. 248].

This last result follows also from [7, p. 510 (4)], since $j'_{\nu 1}>\nu$ when $\nu>0$ [5, p. 246], [7, p. 485 (1)]. This formula, which will be used below, states

$$\frac{dc'}{d\nu}=\frac{2c'}{c'^2-\nu^2}\int_0^\infty (c'^2\cosh 2t-\nu^2)K_0(2c'\sinh t)\,e^{-2\nu t}\,dt, \tag{4}$$

where the positive decreasing function $K_0(x)$ is the modified Bessel function of the second kind and order 0 [7, p. 78].

* Received by the editors July 13, 1987; accepted for publication December 18, 1987. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, York University, North York, Ontario, Canada M3J 1P3.

‡ 75 Glen Eyrie Avenue, San Jose, California 95125.

An alternative proof of the monotonicity of $j'_{\nu k}$, $\nu > 0$, can be based on Sturm comparison techniques instead of (4) [3].

2. Statement of results to be established. If the inequality $j'_{\nu 1} > |\nu|$, known [5, p. 246] to be valid for $\nu > 0$, could be established also for $-1 < \nu < 0$, then (4) would imply also that $j'_{\nu k}$ is an increasing function of ν, $-1 < \nu < 0$, $k = 1, 2, \cdots$. In fact, somewhat more is true, namely,

$$j'_{\nu 1} > \max\{j'_{11}|\nu|, j_{\nu 1}, j'_{-\nu 1}\}, \qquad -1 < \nu < 0, \tag{5}$$

where [6, p. 30] $j'_{11} = 1.84118378 > 3^{1/2}$. Indeed,

$$j'_{\nu 1} > j'_{11}, \qquad -1 < \nu < 0. \tag{5'}$$

Hence, as stated, for $k = 1, 2, \cdots$,

$$j'_{\nu k} \textit{ increases with } \nu, \qquad -1 < \nu < 0, \tag{6}$$

as well as for $\nu > 0$.

This does *not* assert that $j'_{\nu k}$ increases for $-1 < \nu < \infty$. Indeed, such a claim would be false, as the third inequality in (5) shows. The function $j'_{\nu 1}$ is discontinuous at $\nu = 0$, although it is continuous from above at $\nu = 0$, i.e., $j'_{\nu 1} \to j'_{01} = 0$ as $\nu \to 0+$. In the transition from negative ν to positive ν, the zero $j'_{\nu k}$ goes over into $j'_{\nu, k+1}$.

As a consequence of (5) and (3), it will be established that

$$(-1)^k J''_\nu(j'_{\nu k}) > 0, \qquad -1 < \nu < 0. \tag{7}$$

Analogous results hold for $\nu = 0$ and $0 < \nu \leqq 1$, namely,

$$J''_0(j'_{01}) = J''_0(0) = 0, \quad (-1)^k J''_0(j'_{0k}) < 0, \quad k = 2, 3, \cdots, \tag{8}$$

and

$$(-1)^k J''_\nu(j'_{\nu k}) < 0, \quad 0 < \nu \leqq 1, \quad k = 1, 2, \cdots. \tag{9}$$

This last set of inequalities cannot hold for general $\nu > 0$, since they depend on the inequality $j'_{\nu k} > 3^{1/2}\nu$, valid for $0 < \nu \leqq 1$ (in fact for somewhat larger ν as well) but not for sufficiently large ν, as can be noted from the well-known fact that $j'_{\nu 1}/\nu \to 1$ as $\nu \to \infty$.

It will be shown also that

$$c'_\nu - c_\nu \textit{ increases with } \nu, \textit{ provided } c'_\nu > c_\nu > \nu > 0. \tag{10}$$

Here c'_ν and c_ν need not be of the same rank. Special cases of (10) worth noting are, for fixed $k = 1, 2, \cdots$,

$$y'_{\nu k} - y_{\nu k} \textit{ increases with } \nu > 0, \tag{11}$$

and

$$j'_{\nu, k+1} - j_{\nu k} \textit{ increases with } \nu > 0. \tag{12}$$

Remark. The monotonicity result (10) and its corollaries (11) and (12) are reminiscent of the results in [4] where it was shown, e.g., that $j_{\nu, k+1} - j_{\nu k}$ and similar differences are increasing functions of $\nu > 0$ for each fixed $k = 1, 2, \cdots$. Proof was done by employing the integral representation for $dc/d\nu$ which will be used here to establish (10). These and other results were later derived by Sturm techniques in [3].

For the corresponding differences such as $j'_{\nu,k+1}-j'_{\nu k}$, $k=1,2,\cdots$, we have not proved any monotonicity properties. Numerical values suggest these differences and the differences $j_{\nu k}-j'_{\nu k}$ possess regularities which are less uniform. Thus, $j_{\nu 1}-j'_{\nu 1}$ decreases over the values $\nu=0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}$ but increases when $\nu=\frac{3}{4}, 1, \frac{3}{2}, \cdots$.

3. Proof of (5) and (6). As a preliminary to (5), but sufficient to establish (6), we note first

$$j'_{\nu 1}>|\nu|, \qquad -1<\nu<0. \tag{13}$$

This is a consequence of $J'_\nu(x)<0$, $|x|\leqq|\nu|$, $-1<\nu<0$, which in turn follows on differentiating (1) and isolating the first term of the resulting infinite series to obtain

$$2^\nu x^{1-\nu}J'_\nu(x)=\frac{\nu}{\Gamma(\nu+1)}-\sum_{m=1}^{\infty}\frac{(-1)^{m+1}(2m+\nu)x^{2m}}{4^m m!\Gamma(m+\nu+1)}.$$

The isolated term on the right side is negative for $-1<\nu<0$ and the (alternating) infinite series is positive for $0<|x|\leqq|\nu|$, since its successive terms decrease in absolute value for these x. Hence, (13) holds.

This implies also (6), on taking $c'=j'_{\nu k}$ in (4).

Putting $x=j'_{\nu 1}$ in the differential equation (2), with $y=J_\nu(x)$, we now infer also that

$$J_\nu(j'_{\nu 1})<0, \qquad -1<\nu<0, \tag{14}$$

since $x=j'_{\nu 1}$ yields a minimum for $J_\nu(x)$, $-1<\nu<0$. Thus, $j'_{\nu 1}>j_{\nu 1}$, as asserted in the second inequality in (5), since $J_\nu(0+)=+\infty$ for such ν.

To establish the third inequality in (5), we recall the Wronskian [7, p. 76]

$$J_\nu(x)J'_{-\nu}(x)-J'_\nu(x)J_{-\nu}(x)=-(2\sin\nu\pi)/(\pi x). \tag{15}$$

This implies $J_\nu(j'_{\nu 1})J'_{-\nu}(j'_{\nu 1})>0$ when, as here, $-1<\nu<0$, so that $J'_{-\nu}(j'_{\nu 1})<0$, from (14). Hence $j'_{\nu 1}>j'_{-\nu 1}$, as the third inequality in (5) states, since $-\nu>0$.

It remains to verify the first inequality in (5). When we know that $j'_{\nu 1}$ and $j'_{-\nu 1}$ are monotonic for $-1<\nu<0$, it follows that their respective limits exist as $\nu\to-1+$. Since $j'_{\nu 1}>j'_{-\nu 1}$, $-1<\nu<0$, we have

$$\lim_{\nu\to-1+} j'_{\nu 1}\geqq\lim_{\nu\to-1+} j'_{-\nu,1}=j'_{11},$$

in view of the continuity of $j'_{\nu 1}$ at $\nu=1$ [5, p. 246]. The first inequality in (5) now follows, since (6) has already been verified. The reasoning provided establishes also the stronger assertion (5′).

4. Proof of (7), (8), and (9). The first inequality in (5), slightly weakened, implies

$$j'_{\nu k}>3^{1/2}|\nu|, \qquad -1<\nu<0. \tag{16}$$

In differential equation (3), we substitute $x=j'_{\nu k}$, with $y=J_\nu(x)$, and recall that $j'_{\nu 1}, j'_{\nu 3}, \cdots$, yield minima, $j'_{\nu 2}, j'_{\nu 4}, \cdots$, yield maxima when $-1<\nu<0$. The y' term vanishes; the coefficients of y'' and y''' are positive in view of (16), so that (7) follows from (3).

The results (8) and (9) can be shown similarly, except that, for the first part of (8), we start by remembering that $j'_{01}=0$. That $J'''_0(j'_{01})=0$ then follows by differentiating (1) three times and then putting $x=0$. For $0<\nu\leqq 1$, we need to know that $j'_{\nu 1}>\sqrt{3}\nu$. This follows from the inequality [7, p. 486 (3)]

$$j'_{\nu 1}>\sqrt{\nu(\nu+2)}, \qquad \nu>0.$$

Remark. In fact, $j'_{\nu 1} > j'_{11}\nu$, $0 < \nu < 1$. This follows from the concavity of $j'_{\nu k}$, established by Elbert and Laforgia [1], since $j'_{01} = 0$ so that the chord joining the origin with $(1, j'_{11})$ lies below the curve $j'_{\nu 1}$ in the $(\nu, j'_{\nu 1})$ plane.

5. Proof of (10), (11), and (12). These are consequences of (4) and the lemma of [4], since $\cosh 2t > 1$, $t > 0$. Thus, from (4) and [4],

$$\begin{aligned}\frac{dc'_\nu}{d\nu} &= 2c'_\nu \int_0^\infty \frac{c'^2_\nu \cosh 2t - \nu^2}{c'^2_\nu - \nu^2} K_0(2c'_\nu \sinh t)\, e^{-2\nu t}\, dt \\ &> 2c'_\nu \int_0^\infty K_0(2c'_\nu \sinh t)\, e^{-2\nu t}\, dt \\ &> 2c_\nu \int_0^\infty K_0(2c_\nu \sinh t)\, e^{-2\nu t}\, dt,\end{aligned}$$

where the last inequality follows from the lemma of [4] since $c'_\nu > c_\nu > \nu > 0$. But this last expression equals $dc_\nu/d\nu$ [7, p. 508 (3)], verifying (10). The remaining assertions (11) and (12) are corollaries.

6. An additional remark. When we put $x = j'_{\nu 1}$ in the Wronskian

$$W(J_\nu, Y_\nu) = J_\nu(x) Y'_\nu(x) - J'_\nu(x) Y_\nu(x) = 2/(\pi x),$$

it follows that $J_\nu(j'_{\nu 1}) Y'_\nu(j'_{\nu 1}) > 0$. When $-1 < \nu < 0$, the factor $J_\nu(j'_{\nu 1}) < 0$, and so for this interval of ν, $Y'_\nu(j'_{\nu 1}) < 0$. When $-\frac{1}{2} < \nu < 0$, $Y_\nu(0+) = -\infty$, while $Y_{-1/2}(x) = J_{1/2}(x)$, so that $Y_{-1/2}(0) = 0$. Hence

$$y'_{\nu 1} < j'_{\nu 1}, \qquad -\tfrac{1}{2} \leqq \nu < 0, \tag{17}$$

reversing the inequality which obtains for $\nu > 0$ [6, p. xvi (1.04)].

It is perhaps worth noticing the significance of (16) by inferring (17) from another "Wronskian" [7, p. 76 (8)], namely

$$J'_\nu(x) Y'''_\nu(x) - J'''_\nu(x) Y'_\nu(x) = \frac{2}{\pi x^2}\left(\frac{3\nu^2}{x^2} - 1\right). \tag{18}$$

When $x = j'_{\nu 1}$, this implies

$$J'''_\nu(j'_{\nu 1}) Y'_\nu(j'_{\nu 1}) > 0, \qquad -1 < \nu < 0,$$

in view of (16). From (7), it follows now that $Y'_\nu(j'_{\nu 1}) < 0$, as in the previous proof, leading again to (17).

Note added in proof. The zeros of $J'_\nu(x)$ are simple when $\nu > -1$, except possibly for $x = 0$. This follows on applying inequality (13) to the differential equation (2), since (13) holds also when $\nu \geqq 0$. For $\nu < -1$, this is no longer the case. This is pointed out by Kerimov and Skorokhodov, "Calculation of the multiple zeros of the derivatives of the cylindrical Bessel functions $J_\nu(z)$ and $Y_\nu(z)$," *Zh. Vychisl. Mat. i Mat. Fiz.*, 25 (1985), pp. 1749–1760 (*U.S.S.R. Comput. Math. and Math. Phys.*, 25 (1985), pp. 101–107, especially p. 107).

REFERENCES

[1] A. ELBERT AND A. LAFORGIA, *On the zeros of derivatives of Bessel functions*, J. Appl. Math. Phys. (ZAMP), 34 (1983), pp. 774–786.

[2] A. ERDÉLYI, ED., *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, Toronto, London, 1953.

[3] L. LORCH, *Elementary comparison techniques for certain classes of Sturm–Liouville equations*, Proc. Uppsala 1977 International Conference on Differential Equations, Symposia Univ. Upsaliensis Annum Quingentesimum Celebrantis 7, Acta Univ. Upsaliensis, Uppsala, 1977, pp. 125–133.

[4] L. LORCH AND P. SZEGO, *Monotonicity of the difference of zeros of Bessel functions as a function of order*, Proc. Amer. Math. Soc., 15 (1964), pp. 91–96.

[5] F. W. J. OLVER, *Asymptotics and Special Functions*, Academic Press, New York, London, 1974.

[6] F. W. J. OLVER, ED., *Royal Society Mathematical Tables*, Vol. 7, *Bessel Functions. Part* III. *Zeros and Associated Values.* The University Press, Cambridge, 1960.

[7] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., The University Press, Cambridge, 1952.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

018

A COMBINATORIAL INTERPRETATION OF THE INTEGRAL OF THE PRODUCT OF LEGENDRE POLYNOMIALS*

J. GILLIS†, J. JEDWAB‡, AND D. ZEILBERGER§

Abstract. Denote by $P_n(x)$ the Legendre polynomial of degree n and let

$$I_{n_1,\cdots,n_k} = \int_{-1}^{1} P_{n_1}(x)\cdots P_{n_k}(x)\,dx.$$

$I_{n_1,\cdots,n_k}$ is written as a sum involving binomial coefficients and the sum is interpreted via a combinatorial model. This makes possible a combinatorial proof of a number of general theorems concerning $I_{n_1,\cdots,n_k}$, not all of which seem analytically straightforward, including a direct combinatorial derivation of the known formula for $I_{a,b,c}$ and the expression of $I_{a,b,c,d}$ as a simple finite sum. In addition, a number of apparently new combinatorial identities are obtained.

Key words. Legendre polynomials, integrals, digraph

AMS(MOS) subject classifications. primary 33A45; secondary 05A10

1. Introduction.

We will be concerned with the Legendre polynomials, defined by

$$P_n(x) = 2^{-n} \sum_{\alpha \leqq n/2} (-1)^{\alpha} \binom{n}{\alpha} \binom{2n-2\alpha}{n} x^{n-2\alpha} \qquad (-1 \leqq x \leqq 1;\ n = 0, 1, 2, \cdots),$$

which may be written in the equivalent form [4, p. 38]

$$P_n(x) = 2^{-n} \sum_{\alpha} \binom{n}{\alpha}^2 (x+1)^{\alpha}(x-1)^{n-\alpha}. \tag{1}$$

In (1), as in other combinatorial sums in what follows, we shall omit the limits of summation where these coincide with the natural cut-offs implied by the fact that $\binom{a}{b} = 0$ wherever a, b are integers and $b > a > 0$ or $a > 0 > b$.

Let

$$I_{n_1,\cdots,n_k} = \int_{-1}^{1} P_{n_1}(x)\cdots P_{n_k}(x)\,dx, \tag{2}$$

where $n_1, \cdots, n_k$ are nonnegative integers. We will express $I_{n_1,\cdots,n_k}$ as a sum involving binomial coefficients and use a combinatorial interpretation of this sum to derive a number of analytical and combinatorial results.

To simplify notation, write $\underline{n} = (n_1, \cdots, n_k)$ and $\underline{\alpha} = (\alpha_1, \cdots, \alpha_k)$. It is convenient to write $x = 2y - 1$ in (1) to obtain

$$P_n(x) = p_n(y) = \sum_{\alpha} \binom{n}{\alpha}^2 y^{\alpha}(y-1)^{n-\alpha} \qquad (0 \leqq y \leqq 1),$$

which on substitution in (2) gives

$$\begin{aligned} I_{\underline{n}} &= 2 \sum_{\underline{\alpha}} \binom{n_1}{\alpha_1}^2 \cdots \binom{n_k}{\alpha_k}^2 \int_0^1 y^{\Sigma\alpha_i}(y-1)^{\Sigma n_i - \Sigma\alpha_i}\,dy \\ &= 2 \sum_{\underline{\alpha}} \binom{n_1}{\alpha_1}^2 \cdots \binom{n_k}{\alpha_k}^2 \left\{ \frac{(-1)^{\Sigma\alpha_i}(\Sigma\alpha_i)!\,(\Sigma n_i - \Sigma\alpha_i)!}{(1+\Sigma n_i)!} \right\} \end{aligned} \tag{3}$$

* Received by the editors March 25, 1987; accepted for publication (in revised form) March 1, 1988.
† Department of Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel.
‡ Emmanuel College, Cambridge, United Kingdom.
§ Department of Mathematics, Drexel University, Philadelphia, Pennsylvania 19104.

$$(4) \qquad = \frac{2}{1+\Sigma n_i} \cdot \sum_{\alpha} (-1)^{\Sigma \alpha_i} \left(\binom{n_1}{\alpha_1}^2 \cdots \binom{n_k}{\alpha_k}^2 \Big/ \binom{\Sigma n_i}{\Sigma \alpha_i} \right).$$

Now consider a set of elements of k different types, ordered by type number i ($i = 1, \cdots, k$) and, within each type number, by a serial number r ($r = 1, \cdots, n_i$). We represent these by points and form a directed graph by connecting them, with one edge going into and one coming out of each of the points. We then color each of the Σn_i edges blue or yellow according to the following *balance condition*:

(*) For each i the number of points of type i at the beginning of blue edges, α_i (say), equals the number at the end of blue edges.

Call each such colored graph a *system*, and let T denote the set of all possible distinct systems. Class a system as *even* or *odd* according to the parity of the total number of blue edges, $\sum_{i=1}^{k} \alpha_i$. Let the difference between the number of even and odd systems, in any subset E of T, be $\Pi_n(E)$. Clearly,

$$(5) \qquad \begin{aligned} \frac{2}{(1+\Sigma n_i)!} \Pi_n(T) &= \frac{2}{(1+\Sigma n_i)!} \left\{ \sum_{\alpha} (-1)^{\Sigma \alpha_i} \binom{n_1}{\alpha_1}^2 \cdots \binom{n_k}{\alpha_k}^2 (\Sigma \alpha_i)! (\Sigma n_i - \Sigma \alpha_i)! \right\} \\ &= I_n \end{aligned}$$

by (3).

2. Some elementary considerations. Denote the set of distinct graphs formed by omitting the coloring of each system, in any subset E of T, by E^*. We will refer to an edge beginning at a point of type i and ending at a point of type j ($1 \leqq i, j \leqq k$) as "an $i \to j$ edge," calling it *pure* if $i = j$ and *mixed* if $i \neq j$. Where desired we will indicate the edge color by $i \overset{B}{\to} j$ or $i \overset{Y}{\to} j$.

We recall that any vertex is characterized by a pair of natural numbers (i, j) where i is the number of the type to which it belongs and j ($1 \leqq j \leqq n_i$) is its serial number in that type. If two points P, P' are characterized by (i, j), (i', j'), respectively, then P is said to be of *lower rank* than P' if

$$\text{either} \quad i < i'$$

$$\text{or} \quad i = i', \quad j < j'.$$

Now let P be the set of systems containing at least one pure edge. Given any system in P, select from among the pure edges the one beginning at the point of lowest rank and change its color. This leaves the balance condition (*) satisfied but produces a new system of opposite parity, so that the two systems together give a canceling contribution to $\Pi_n(T)$. Since this process defines a (1, 1) parity-changing map from P to itself, $\Pi_n(P) = 0$. Writing $T \backslash P = U$, say, this is equivalent to

$$\Pi_n(T) = \Pi_n(U);$$

thus, we may disregard P and count only the contribution of systems in U to $\Pi_n(T)$.

Now consider any graph belonging to U. Take the lowest ranking vertex and call it X^0. Since there is exactly one edge starting at each vertex, there will be a uniquely defined cycle of the form

$$(6) \qquad X^0 \to X^1 \to X^2 \to \cdots \to X^{P-1} \to X^P (= X^0).$$

Since there are no pure edges it follows that $P \geqq 2$ and that each edge $X^i \to X^{i+1}$ is mixed. If this cycle does not cover the entire graph, let Y^0 be the lowest ranking vertex

not lying on it. As before, we define a cycle

$$Y^0 \to Y^1 \to \cdots \to Y^{q-1} \to Y^q (= Y^0) \tag{7}$$

and continue until the entire graph has been covered in this way.

Consider any such cycle, e.g., (6). It can be divided up into segments each of which begins and ends with vertices of the same *type* as X^0. For example, if the cycle (6) were

$$(1,1) \to (2,5) \to (3,7) \to (1,6) \to (2,4) \to (7,1) \to (4,3) \to (1,2) \to (2,1) \to (1,1),$$

we should have the segments

$$(1,1) \to (2,5) \to (3,7) \to (1,6),$$

$$(1,6) \to (2,4) \to (7,1) \to (4,3) \to (1,2),$$

$$(1,2) \to (2,1) \to (1,1).$$

We can thus describe the graph structure by a set of segments, which we may order by the ranks of their initial vertices. A segment will be called *odd* or *even* according to the parity of the number of edges which compose it. Now let V be the set of graphs whose structures contain at least one odd segment. Suppose such a graph, G (say), contains the segment

$$Z^0 \to Z^1 \to \cdots \to Z^{r-1} \to Z^r \tag{8}$$

where r is odd, and suppose, moreover, that (8) is the lowest ranking such odd segment in this graph. We change the graph, and its coloring, according to the following rules:

(a) Change the connecting edges of (8) to produce the segment

$$Z^0 \to Z^{r-1} \to Z^{r-2} \to \cdots \to Z^1 \to Z^r. \tag{9}$$

(b) Color these new edges so that the ith edge of (9) ($1 \leqq i \leqq r$) has the opposite color to that of the $(r+1-i)$th edge of (8).

Call the new system, with its coloring, G'. It is easily verified that G' also satisfies the rule (*). On the other hand, since r is odd, the graphs G, G' will have opposite parities. Moreover the transformation is clearly an involution. We thus have a $(1,1)$ parity changing surjection of V onto itself. It follows that

$$\Pi_{\underline{n}}(V) = 0. \tag{10}$$

Therefore if we write $W = U \setminus V$ we have

$$\Pi_{\underline{n}}(U) = \Pi_{\underline{n}}(W), \tag{11}$$

and therefore it is sufficient to construct the systems belonging to W and calculate their contribution to $\Pi_{\underline{n}}(T)$.

3. Application to the case $k=3$. We now apply these considerations to the particular case $k=3$, writing $\underline{n} = (a, b, c)$. Since the product $P_a(x)P_b(x)P_c(x)$ is a polynomial of parity equal to that of $a+b+c$, its integral will be zero for odd $a+b+c$. We therefore limit the discussion to $a+b+c = 2s$, where s is an integer. Moreover it follows from the orthogonality of the polynomials that the integral will vanish unless $s \geqq \max(a, b, c)$. We proceed to study the integral under these assumptions. Let G be any graph of the set W^*, let E be the number of even systems that can be constructed

by coloring G, and let Ω be the number of odd systems. By the definition of W, each such graph is made up of segments of one of the following forms:

$$\begin{aligned} &\text{(i)} \quad (2\to 3\to)^{l_i}2, \\ &\text{(ii)} \quad 1\to(3\to 2\to)^{m_i}3\to 1, \\ &\text{(iii)} \quad 1\to(2\to 3\to)^{n_i}2\to 1, \end{aligned} \tag{12}$$

where the 1, 2, 3 indicate the types to which the vertices belong and $l_i \geqq 1$, $m_i \geqq 0$, $n_i \geqq 0$. Let the segments in each of these three classes be ordered by the rank of their initial vertex.

For each i, j $(i, j = 1, 2, 3)$ denote by E_{ij} the number of $i \to j$ edges. In any segment of type (i), (ii), or (iii) the number of $i \to j$ edges equals that of $j \to i$ edges and, hence, for the whole graph $E_{ij} = E_{ji}$. Since, by hypothesis, there are no pure edges, it follows that

$$\begin{aligned} &E_{12} = E_{21} = s - c, \\ &E_{13} = E_{31} = s - b, \quad \text{and} \\ &E_{23} = E_{32} = s - a. \end{aligned} \tag{13}$$

To simplify the notation we write A, B, C for $s-a, s-b, s-c$, respectively. It is easily seen that the only possible distributions of colors consistent with (*) must be as shown in the following table:

	$2\to 3$	$3\to 2$	$3\to 1$	$1\to 3$	$1\to 2$	$2\to 1$
Blue	$\alpha + t$	α	$\beta + t$	β	$\gamma + t$	γ
Yellow	$A - \alpha - t$	$A - \alpha$	$B - \beta - t$	$B - \beta$	$C - \gamma - t$	$C - \gamma$

where t, α, β, γ may take any values for which the table entries are all nonnegative integers. Now the distribution of colors among the $2\to 3$ and $3\to 2$ edges is determined when we have chosen $(\alpha + t)$ of the $2\to 3$ edges and $A - \alpha$ from the $3\to 2$ edges, and this can clearly be done in $\binom{2A}{A+t}$ ways. Similar results hold for $3\to 1$ and for $1\to 2$. The total number of blue edges in any such coloring is $2(\alpha+\beta+\gamma)+3t \equiv t \pmod 2$. Hence the difference between the numbers of even and odd systems possible on any such graph is

$$\begin{aligned} &\sum_t (-1)^t \binom{2A}{A+t}\binom{2B}{B+t}\binom{2C}{C+t} \\ &\quad = \frac{(2A)!(2B)!(2C)!}{(B+C)!(C+A)!(A+B)!} \sum_t (-1)^t \binom{A+B}{A+t}\binom{B+C}{B+t}\binom{C+A}{C+t}. \end{aligned} \tag{14}$$

But

$$\sum_t (-1)^t \binom{A+B}{A+t}\binom{B+C}{B+t}\binom{C+A}{C+t} = \frac{(A+B+C)!}{A!B!C!}. \tag{15}$$

(For an elegant combinatorial proof of this known identity, see [3, p. 65].) Substituting into (17), we obtain

$$\sum_t (-1)^t \binom{2A}{A+t}\binom{2B}{B+t}\binom{2C}{C+t}$$

$$
\begin{aligned}
&= \frac{(2A)!(2B)!(2C)!}{(B+C)!(C+A)!(A+B)!} \cdot \frac{(A+B+C)!}{A!B!C!} \\
&= \frac{(2s-2a)!(2s-2b)!(2s-2c)!s!}{a!b!c!(s-a)!(s-b)!(s-c)!} \\
&= \binom{2s-2a}{s-a}\binom{2s-2b}{s-b}\binom{2s-2c}{s-c} \cdot \frac{s!(s-a)!(s-b)!(s-c)!}{a!b!c!}.
\end{aligned} \tag{16}
$$

In particular the number is the same for all the graphs of W^*. It remains to determine how many such graphs there are.

Consider (12). Since there are B edges of type $1 \to 3$ in the graph, this will also be the number of segments of type (ii). Similarly there will be C segments of type (iii). Hence the total number of m_i's and n_i's is $B+C=a$. If we write $L=\Sigma l_i$ we see we have to determine the numbers $L, \{m_i\}, \{n_i\}$. Since $L+\Sigma m_i+\Sigma n_j$ equals the number of $2 \to 3$ edges, i.e., A, it follows that $L, \{m_i\}, \{n_j\}$ are the nonnegative integer solutions of

$$\sum_{i=1}^{a+1} x_i = A$$

and this number is known to be $\binom{A+a}{a} = \binom{s}{a}$.

The number of possibilities with the segments in each of (ii) and (iii) ranked in order is therefore $\binom{s}{a}/(B!C!)$. If the segments of forms (ii), (iii) are connected via vertices of type 1 (and this may be done in $a!$ ways), the graph will be determined except for the numbers l_i and the ranks of the vertices. The vertices not involved in segments of form (i) may be ranked in $a!(b!/L!)(C!/L!)$ ways while it is easily seen that the remaining pairs of $(1, 2)$ points may be connected in cycles and ranked in $(L!)^2$ ways. Hence the total number of possible graphs in W^* is

$$\left\{\binom{s}{a}\Big/(B!C!)\right\} \cdot a! \cdot \{a!b!c!\} = \frac{s!a!b!c!}{A!B!C!}. \tag{17}$$

It follows from (5), (16), and (17) that

$$I_{a,b,c} = \frac{2}{(a+b+c+1)} \cdot \binom{2s-2a}{s-a}\binom{2s-2b}{s-b}\binom{2s-2c}{s-c}\binom{2s}{s}^{-1}. \tag{18}$$

This result was first obtained by Adams [1]. His approach was to evaluate the integral for some low values of the subscripts and, on the basis of this, to guess a general formula, which he then proved by induction. For a succinct history of the problem see Askey [2, pp. 39–40]. In the special case $a=b=c=2\lambda$, (18) becomes

$$\int_{-1}^{1} \{P_{2\lambda}(x)\}^3 \, dx = \frac{2\{(3\lambda)!\}^2}{(6\lambda+1)!}\binom{2\lambda}{\lambda}^3. \tag{19}$$

Substituting (18) into (4), we get the binomial identity

$$
\begin{aligned}
&\sum_{\alpha,\beta,\gamma} (-1)^{\alpha+\beta+\gamma}\binom{a}{\alpha}^2\binom{b}{\beta}^2\binom{c}{\gamma}^2\binom{a+b+c}{\alpha+\beta+\gamma}^{-1} \\
&= \begin{cases} \binom{2s-2a}{s-a}\binom{2s-2b}{s-b}\binom{2s-2c}{s-c}\binom{2s}{s}^{-1} & \text{for } a+b+c=2s, \\ 0 & \text{for } a+b+c \text{ odd.} \end{cases}
\end{aligned} \tag{20}
$$

In the special case $a=b=c=2\lambda$, this becomes

$$\sum_{\alpha,\beta,\gamma} (-1)^{\alpha+\beta+\gamma}\binom{2\lambda}{\alpha}^2\binom{2\lambda}{\beta}^2\binom{2\lambda}{\gamma}^2\binom{6\lambda}{\alpha+\beta+\gamma}^{-1} = \binom{2\lambda}{\lambda}^3\binom{6\lambda}{3\lambda}^{-1}. \tag{21}$$

4. The case $k=4$. Since the product $P_{n_1}, P_{n_2} \cdots P_{n_k}$ is a polynomial, it may be written in the form

$$P_{n_1}(x)P_{n_2}(x)\cdots P_{n_k}(x)=\sum_{\alpha} C_{n_1,n_2,\cdots,n_k,\alpha}P_\alpha(x) \tag{22}$$

where the $C_{n_1,\cdots,n_k}$, α are constants. Now if we apply the well-known relation

$$\int_{-1}^{1} P_m(x)P_n(x)\,dx=\frac{2}{2m+1}\delta_{m,n}, \tag{23}$$

we get

$$\frac{2}{2\alpha+1}C_{n_1,n_2,\cdots,n_k,\alpha}=I_{n_1,n_2,\cdots,n_k,\alpha}. \tag{24}$$

Now let a, b, c, d be nonnegative integers. By (23) and (24)

$$\begin{aligned} P_a(x)P_b(x)&=\sum_{\alpha}\left(\alpha+\frac{1}{2}\right)I_{a,b,\alpha}P_\alpha(x), \quad \text{and} \\ P_c(x)P_d(x)&=\sum_{\beta}\left(\beta+\frac{1}{2}\right)I_{c,d,\beta}P_\beta(x). \end{aligned} \tag{25}$$

Hence,

$$\begin{aligned} I_{a,b,c,d}&=\int_{-1}^{1} P_a(x)P_b(x)P_c(x)P_d(x)\,dx \\ &=\sum_{\alpha,\beta}\left(\alpha+\frac{1}{2}\right)\left(\beta+\frac{1}{2}\right)I_{a,b,\alpha}I_{c,d,\beta}\int_{-1}^{1}P_\alpha(x)P_\beta(x)\,dx \\ &=\sum_{\alpha,\beta}\left(\alpha+\frac{1}{2}\right)\left(\beta+\frac{1}{2}\right)I_{a,b,\alpha}I_{c,d,\beta}\frac{\delta_{\alpha,\beta}}{\alpha+\frac{1}{2}} \quad \text{by (22)} \\ &=\sum_{\alpha}\left(\alpha+\frac{1}{2}\right)I_{a,b,\alpha}I_{c,d,\alpha}. \end{aligned} \tag{26}$$

Since $I_{a,b,c,d}=0$, unless $a+b+c+d$ is even, we may assume in (26) that $a+b\equiv c+d \pmod 2$. Thus we may write

$$\begin{aligned} I_{a,b,c,d}&=\sum_{\gamma}\left(2\gamma+\frac{1}{2}\right)I_{a,b,2\gamma}I_{c,d,2\gamma} \quad \text{if } a+b\equiv c+d\equiv 0 \pmod 2 \\ &=\sum_{\gamma}\left(2\gamma+\frac{3}{2}\right)I_{a,b,2\gamma+1}I_{c,d,2\gamma+1} \quad \text{if } a+b\equiv c+d\equiv 1 \pmod 2. \end{aligned} \tag{27}$$

Moreover, since $I_{a,b,c,d}$ is clearly symmetric in the subscripts, we see that

$$\sum_{\alpha}\left(\alpha+\frac{1}{2}\right)I_{a,b,\alpha}I_{c,d,\alpha}=\sum_{\beta}\left(\beta+\frac{1}{2}\right)I_{a,c,\beta}I_{b,d,\beta}. \tag{28}$$

A special case of some interest arises if we take $a=b=c=d$. By (27) we get

$$I_{a,a,a,a}=\sum_{\gamma}\left(2\gamma+\frac{1}{2}\right)I_{a,a,2\gamma}^2,$$

i.e.,

$$\int_{-1}^{+1}[P_a(x)]^4\,dx=2\sum_{\gamma}(4\gamma+1)\left\{\frac{[(a+\gamma)!]^2}{(2a+2\gamma+1)}\binom{2\gamma}{\gamma}^2\binom{2a-2\gamma}{a-\gamma}\right\}^2. \tag{29}$$

REFERENCES

[1] J. C. ADAMS, Proc. Roy. Soc., XXVII (1878).
[2] R. ASKEY, *Orthogonal Polynomials and Special Functions*, CBMS-NSF Regional Conference Series in Applied Mathematics 21, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.
[3] P. CARTIER AND D. FOATA, *Problèmes combinatoires de commutation et de rearrangement*, Lecture Notes in Math. 85, Springer-Verlag, Berlin, New York, 1969.
[4] H. W. GOULD, *Combinatorial Identities*, 2nd ed., University of West Virginia, Morgantown, WV, 1972.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

019

A BETA INTEGRAL ASSOCIATED WITH THE ROOT SYSTEM G_2*

F. G. GARVAN†

Abstract. Some conjectures of Askey are proven that have to do with adding roots in the Macdonald-Morris conjecture for G_2. This is done by extending Aomoto's proof of Selberg's integral. This yields a new proof of the Macdonald-Morris root system conjecture for G_2 which should extend to other root systems.

Key words. Askey's G_2 conjectures, Macdonald-Morris root system conjectures, Aomoto, Selberg's integral, multidimensional beta integrals

AMS(MOS) subject classifications. primary 33A15, 33A75

1. Introduction.

Let

(1.1)

$$\begin{aligned} G(x_1,x_2;a,b) &= (1-x_1)^a\left(1-\frac{1}{x_1}\right)^a(1-x_2)^a\left(1-\frac{1}{x_2}\right)^a(1-x_1x_2)^a\left(1-\frac{1}{x_1x_2}\right)^a \\ &\cdot\left(1-\frac{x_1}{x_2}\right)^b\left(1-\frac{x_2}{x_1}\right)^b(1-x_1^2x_2)^b\left(1-\frac{1}{x_1^2x_2}\right)^b(1-x_2^2x_1)^b\left(1-\frac{1}{x_2^2x_1}\right)^b. \end{aligned}$$

Then the Macdonald-Morris root system conjecture for G_2 is

$$\text{(1.2)}\qquad \text{C.T. } G(x_1,x_2;a,b)=\frac{(3a+3b)!(3b)!(2a)!(2b)!}{(2a+3b)!(a+2b)!(a+b)!a!b!b!}=g(a,b).$$

Here C.T. means the constant term in the Laurent expansion as a polynomial in x_1, x_1^{-1}, x_2, x_2^{-1}. This has been proved independently by Habsieger [5] and Zeilberger [12]. They have also proved the q-analogue of (1.2). Although their proofs are elegant, they are special to G_2. Recently Zeilberger [13] has also proved the $G_2^\vee$ case of the Macdonald-Morris conjectures. His proof should extend to other root systems.

In this paper we give a new proof of (1.2) which is entirely in terms of integrals and which should also extend to other root systems. Our proof is an extension of Aomoto's [1] proof of Selberg's [11] integral. See Askey [3] for a good exposition of Aomoto's proof. We were led to this by considering conjectures of Askey [3] that have to do with adding roots in the Macdonald-Morris root system conjecture for G_2. Askey conjectured

$$\text{(1.3)}\qquad \text{C.T. } (1-x_1)\left(1-\frac{1}{x_1}\right)G(x_1,x_2;a,b)=\frac{2(3a+3b+1)}{2a+3b+1}g(a,b),$$

$$\text{(1.4)}\qquad \text{C.T. } (1-x_1)\left(1-\frac{1}{x_1}\right)(1-x_2)\left(1-\frac{1}{x_2}\right)G(x_1,x_2;a,b)$$
$$=\frac{2(3a+3b+2)(3a+3b+1)}{(2a+3b+1)(a+2b+1)}g(a,b),$$

*Received by the editors October 5, 1987; accepted for publication February 21, 1988.

†Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706. Current address, School of Mathematics, Macquarie University, Sydney, New South Wales 2109, Australia.

(1.5)

$$\text{C.T. } (1-x_1^2x_2)\left(1-\frac{1}{x_1^2x_2}\right)G(x_1,x_2;a,b)=\frac{2(3a+3b+1)(3b+1)}{(2a+3b+1)(a+2b+1)}g(a,b),$$

(1.6)

$$\text{C.T. } (1-x_1^2x_2)\left(1-\frac{1}{x_1^2x_2}\right)(1-x_2^2x_1)\left(1-\frac{1}{x_2^2x_1}\right)G(x_1,x_2;a,b)$$
$$=\frac{6(3a+3b+2)(3a+3b+1)(3b+1)(3b+2)}{(2a+3b+3)(2a+3b+2)(2a+3b+1)(a+2b+1)}g(a,b).$$

In §2 we give the idea of the proof of (1.2). It is well known that (1.2) may be written as a trigonometric integral formula (see, for example, Morris [10, p.46]). The starting point of our proof is to write (1.2) as

$$\frac{4^{3a+3b}}{\pi^2}\int_{\mathbb{R}^2}\frac{(t_1+t_2)^{2a}(t_1-t_2)^{2b}(t_2^2+2t_1t_2-1)^{2b}(t_1^2+2t_1t_2-1)^{2b}}{(1+t_1^2)^{2a+4b+1}(1+t_2^2)^{2a+4b+1}}dt_1dt_2 \tag{1.7}$$
$$=g(a,b).$$

The left-hand side of (1.7) is the integral referred to in the title of this paper. This is done via three changes of variables: first, by letting $x_j=e^{2i\theta_j}$ $(j=1,2)$ in (1.1) and using the orthogonality of the exponentials on $[0,\pi]$ to obtain an integral on $[0,\pi]^2$; second, by linear change of variables to obtain an integral on $[-\frac{\pi}{2},\frac{\pi}{2}]^2$; and finally by letting $t_j=\tan\theta_j$ to obtain the integral over $\mathbb{R}^2$. We note that this is the same change of variables that Morris used in transforming his constant terms formula for A_n [10, p. 95] into the Cauchy-Selberg integral [10, (6.6)]. The advantage of this integral over the trigonometric integrals is that the integrand is a rational function of t_1,t_2, which can be easily manipulated using a computer algebra package like REDUCE. In §3 we prove some preliminary results. In §4 we complete the proof of (1.2) and prove (1.3)–(1.5) as well.

Macdonald [8, conjecture(6.1)] has also conjectured generalizations of Mehta's integral formula for arbitary root systems [9], [8, (4.1)]. The G_2 case does not seem to be related to our integral given in (1.7). The Macdonald-Mehta integral conjecture involves parameters that are constant on root length. We note that the two parameter case of the G_2 Macdonald-Mehta integral may be proved in the same way as the one parameter case, which was proved by Macdonald [8, p.1002]. This is done by transforming to polar coordinates. The resulting integral turns out to be the product of a gamma integral and a beta integral.

We should mention that (1.3)–(1.6) and their q-analogues may be proved by other methods. Zeilberger [12] proved (1.2) using the result of Morris, mentioned above, related to A_n and Dixon's [4, §3.1] summation of a well-poised ${}_3F_2$. Equations (1.3)–(1.6) could be proved by trying to generalize Morris's results and Dixon's summation. See Kadell [6] for such generalizations of q-analogues of Morris's results and Askey [2] for some extensions of Dixon's summation. The author has proved (1.3)–(1.6), as well as other similar results, by these methods. In §4 we state these other results without proof. For the most part we restrict attention to (1.2)–(1.5), preferring to take a more direct approach.

Kadell [7] has found yet another approach to proving (1.2) which should extend to other root systems. This involves working with the function $G(x_1,x_2;a,b)$ directly rather than writing it as an integral, and using the fact that derivatives have no residues. Finally, (1.2)–(1.6) could be proved by extending Zeilberger's [13] method for the $G_2^\vee$ case and then letting $q\to 1$.

2. The idea of the proof. We label the roots of G_2 as in Fig. 1.

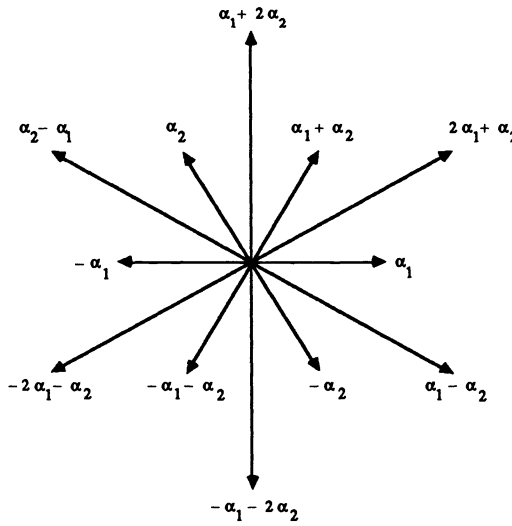


FIG. 1

Let

$$g'(a,b) = \text{C.T. } G(x_1,x_2;a,b). \tag{2.1}$$

Our goal is to prove that $g'(a,b) = g(a,b)$ for all $a,b \geq 0$. The idea is to proceed by induction on a. However to jump from a to $a+1$ in one step would be too much to ask for. Considering (1.3) and (1.4) we break it up into three stages:

Stage 1. $$\text{C.T. } [\alpha_1]G = \frac{2(3a+3b+1)}{2a+3b+1}g'(a,b).$$

Stage 2. $$\text{C.T. } [\alpha_1][\alpha_2]G = \frac{2(3a+3b+1)(3a+3b+2)}{(2a+3b+1)(a+2b+1)}g'(a,b).$$

Stage 3.

$$\text{C.T. } [\alpha_1][\alpha_2][\alpha_1+\alpha_2]G = \frac{6(3a+3b+2)(3a+3b+1)(2a+1)}{(2a+3b+2)(2a+3b+1)(a+2b+1)}g'(a,b)$$

where

$$[k_1\alpha_1+k_2\alpha_2] = (1-x_1^{k_1}x_2^{k_2})(1-x_1^{-k_1}x_2^{-k_2}),$$

for $k_1,k_2 \in \mathbb{Z}$. Each stage corresponds to adding an additional pair of opposite short roots. After Stage 3 all that will remain is to prove the result for $a=0$ since

$$\frac{g(a+1,b)}{g(a,b)} = \frac{6(3a+3b+2)(3a+3b+1)(2a+1)}{(2a+3b+2)(2a+3b+1)(a+2b+1)}. \tag{2.2}$$

The case $a=0$ is equivalent to $b=0$ since we have

$$\text{Long roots of } G_2 \cong \text{Short roots of } G_2 \cong A_2.$$

The result is trivially true for $a=b=0$. The case $b=0$ follows from (2.2) and Stage 3 by induction.

To prove Stages 1–3 we rewrite each stage as an integral and use an idea of Aomoto. To give the reader a taste of our method we work through the proof of Stage 1. If we let

(2.3)
$$w(\underset{\sim}{t}) = w(t_1, t_2; a, b) = \frac{(t_1+t_2)^{2a}(t_1-t_2)^{2b}(t_2^2+2t_1t_2-1)^{2b}(t_1^2+2t_1t_2-1)^{2b}}{(1+t_1^2)^{2a+4b+1}(1+t_2^2)^{2a+4b+1}},$$

then we may write Stage 1 as

$$\frac{4^{3a+3b}}{\pi^2}\int_{\mathbf{R}^2}\frac{4}{(1+t_1^2)}w(\underset{\sim}{t})\,d\underset{\sim}{t} = \frac{2(3a+3b+1)}{(2a+3b+1)}\frac{4^{3a+3b}}{\pi^2}\int_{\mathbf{R}^2}w(\underset{\sim}{t})\,d\underset{\sim}{t}, \tag{2.4}$$

using the same change of variables used to derive (1.7). It is here that we use Aomoto's idea. To get $\int_{\mathbf{R}^2}\frac{1}{1+t_1^2}w(\underset{\sim}{t})\,d\underset{\sim}{t}$ in terms of $\int_{\mathbf{R}^2}w(\underset{\sim}{t})\,d\underset{\sim}{t}$ we use

$$\begin{aligned}0 &= \int_{\mathbf{R}^2}\frac{\partial}{\partial t_1}t_1w(\underset{\sim}{t})\,d\underset{\sim}{t} \\ &= \int_{\mathbf{R}^2}w(\underset{\sim}{t})\,d\underset{\sim}{t} - (2a+4b+1)\int_{\mathbf{R}^2}\frac{2t_1^2}{1+t_1^2}w(\underset{\sim}{t})\,d\underset{\sim}{t} + 2a\int_{\mathbf{R}^2}\frac{t_1}{t_1+t_2}w(\underset{\sim}{t})\,d\underset{\sim}{t} \\ &\quad + 2b\int_{\mathbf{R}^2}\frac{t_1}{t_1-t_2}w(\underset{\sim}{t})\,d\underset{\sim}{t} + 2b\int_{\mathbf{R}^2}\frac{2t_1t_2}{t_2^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t} \\ &\quad + 2b\int_{\mathbf{R}^2}\frac{2t_1(t_1+t_2)}{t_1^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t}.\end{aligned} \tag{2.5}$$

We can make some progress by using the fact that $w(\underset{\sim}{t})$ is invariant under the transposition $t_1 \leftrightarrow t_2$.

$$\frac{t_1}{t_1 \pm t_2} = 1 \mp \frac{t_2}{t_1 \pm t_2}. \tag{2.6}$$

It follows that

$$\int_{\mathbf{R}^2}\frac{t_1}{t_1 \pm t_2}w(\underset{\sim}{t})\,d\underset{\sim}{t} = \frac{1}{2}\int_{\mathbf{R}^2}w(\underset{\sim}{t})\,d\underset{\sim}{t}. \tag{2.7}$$

$$\begin{aligned}&\int_{\mathbf{R}^2}\frac{t_1t_2}{t_2^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t} + \int_{\mathbf{R}^2}\frac{t_1(t_1+t_2)}{t_1^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t} \\ &\quad = \int_{\mathbf{R}^2}\frac{t_1t_2}{t_1^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t} + \int_{\mathbf{R}^2}\frac{t_1(t_1+t_2)}{t_1^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t} \\ &\qquad\qquad \text{(via } t_1 \leftrightarrow t_2 \text{ on the first integral)} \\ &\quad = \int_{\mathbf{R}^2}w(\underset{\sim}{t})\,d\underset{\sim}{t} + \int_{\mathbf{R}^2}\frac{1}{t_1^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t}.\end{aligned} \tag{2.8}$$

Then (2.5) becomes

$$\begin{aligned}0 &= (-3a-3b-1)\int_{\mathbf{R}^2}w(\underset{\sim}{t})\,d\underset{\sim}{t} + 2(2a+4b+1)\int_{\mathbf{R}^2}\frac{1}{1+t_1^2}w(\underset{\sim}{t})\,d\underset{\sim}{t} \\ &\quad + 4b\int_{\mathbf{R}^2}\frac{1}{t_1^2+2t_1t_2-1}w(\underset{\sim}{t})\,d\underset{\sim}{t}.\end{aligned} \tag{2.9}$$

The problem that remains is to get $\int_{\mathbf{R}^2} \frac{1}{t_1^2+2t_1t_2-1}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$ in terms of $\int_{\mathbf{R}^2} w(\underset{\sim}{t})\, d\underset{\sim}{t}$ and $\int_{\mathbf{R}^2} \frac{1}{1+t_1^2}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$. The transformation $t_1 \leftrightarrow t_2$ is not helpful here. We need another transformation that leaves $w(\underset{\sim}{t})\, d\underset{\sim}{t}$ invariant. The real reason why $w(\underset{\sim}{t})$ is invariant under $t_1 \leftrightarrow t_2$ is that the root system G_2 is invariant under the linear transformation given by

$$\alpha_1 \longmapsto \alpha_2 \qquad \text{and} \qquad \alpha_2 \longmapsto \alpha_1,$$

which is also the reflection through the plane orthogonal to the vector $\alpha_2 - \alpha_1$. Recall that a root system is invariant under any element of the Weyl group, the group generated by the w_α, where α is a root and w_α is the reflection through the hyperplane orthogonal to α. The Weyl group for G_2 is generated by $w_{\alpha_2-\alpha_1}$ and w_{α_1}. The extra integral transformation that we need will correspond to w_{α_1}. We study the action of this reflection on the roots:

$$\begin{aligned} &\alpha_1 \leftrightarrow -\alpha_1, && \alpha_2 \leftrightarrow \alpha_1 + \alpha_2, \\ &\alpha_2 - \alpha_1 \leftrightarrow 2\alpha_1 + \alpha_2, && \alpha_1 + 2\alpha_2 \leftrightarrow \alpha_1 + 2\alpha_2. \end{aligned}$$

We need a transformation

$$f : \mathbf{R}^2 \longrightarrow \mathbf{R}^2, \qquad (t_1, t_2) \longmapsto (f_1(t_1, t_2), f_2(t_1, t_2)) \tag{2.10}$$

with the following action:

$$\begin{gathered} \frac{1}{1+t_1^2} \leftrightarrow \frac{1}{1+t_1^2}, \quad \frac{1}{1+t_2^2} \leftrightarrow \frac{(t_1+t_2)^2}{(1+t_1^2)(1+t_2)^2}, \\ \frac{(t_1-t_2)^2}{(1+t_1^2)(1+t_2^2)} \leftrightarrow \frac{(t_1^2+2t_1t_2-1)^2}{(1+t_1^2)^2(1+t_2^2)}, \quad \frac{(t_2^2+2t_1t_2-1)^2}{(1+t_1^2)(1+t_2^2)^2} \leftrightarrow \frac{(t_2^2+2t_1t_2-1)^2}{(1+t_1^2)(1+t_2^2)^2}. \end{gathered} \tag{2.11}$$

The transformation that does the job is

$$f_1(t_1, t_2) = t_1, \qquad f_2(t_1, t_2) = \frac{1 - t_1t_2}{t_1 + t_2} \qquad (t_1 \neq -t_2). \tag{2.12}$$

In §3 we show that

$$\int_{\mathbf{R}^2} g(\underset{\sim}{t}) w(\underset{\sim}{t})\, d\underset{\sim}{t} = \int_{\mathbf{R}^2} g(f(\underset{\sim}{t})) w(\underset{\sim}{t})\, d\underset{\sim}{t}, \tag{2.13}$$

for a certain restricted class of functions g.

Now we return to the problem of evaluating $\int_{\mathbf{R}^2} \frac{1}{t_1^2+2t_1t_2-1}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$. A routine calculation shows that

$$\frac{1}{t_1^2 + 2t_1t_2 - 1} \xrightarrow{f} \frac{(t_1+t_2)}{(1+t_1^2)(t_1-t_2)}, \tag{2.14}$$

$$\frac{(t_1+t_2)}{(1+t_1^2)(t_1-t_2)} = -\frac{(t_1+t_2)^2}{(1+t_1^2)(1+t_2^2)} + \frac{(t_1+t_2)}{(1+t_2^2)(t_1-t_2)}. \tag{2.15}$$

It follows that

(2.16)

$$\int_{\mathbf{R}^2} \frac{1}{t_1^2 + 2t_1t_2 - 1}\, w(\underset{\sim}{t})\, d\underset{\sim}{t} = \int_{\mathbf{R}^2} \frac{(t_1+t_2)}{(1+t_1^2)(t_1-t_2)}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$$

(by (2.13) and (2.14))

$$= -\frac{1}{2}\int_{\mathbf{R}^2} \frac{(t_1+t_2)^2}{(1+t_1^2)(1+t_2^2)}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$$

(by applying $t_1 \leftrightarrow t_2$ and using (2.15))

$$= -\frac{1}{2}\int_{\mathbf{R}^2} \frac{1}{1+t_1^2}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$$

(by applying f, then $t_1 \leftrightarrow t_2$).

Substituting this into (2.9) gives

$$(3a+3b+1)\int_{\mathbf{R}^2} w(\underset{\sim}{t})\, d\underset{\sim}{t} = 2(2a+3b+1)\int_{\mathbf{R}^2} \frac{1}{1+t_1^2}\, w(\underset{\sim}{t})\, d\underset{\sim}{t}$$

and Stage 1 follows. The proof of Stages 2–3 is analogous and will be given in §4.

3. Preliminary results. The main result of this section is Lemma 3.6. It contains a list of integrals that we will need in the proof of Stages 2–3. In the proof of this lemma we will use the transformation formula (2.13) for f. A more formal statement of this formula is given in Lemma 3.2. The idea of the proof of Lemma 3.2 is to write both sides of (2.13) as an integral over $[0,\pi]^2$, using the same change of variables mentioned after (1.7) in the introduction, use the transformation $T(\theta_1,\theta_2) = (-\theta_1, \theta_1+\theta_2)$ and apply Lemma 3.1. The proofs of Lemmas 3.1 and 3.2 are omitted. In the proof of Lemma 3.6 we will also need to calculate the image of certain rational functions in t_1, t_2 under f. This was done using the computer algebra package REDUCE.

LEMMA 3.1. *Let $h : [0,\pi]^2 \longrightarrow \mathbb{R}$ be continuous; then*

$$\int_{[0,\pi]^2} h(\theta_1,\theta_2)\, d\theta_1 d\theta_2 = \int_{[0,\pi]^2} h^*(\theta_1,\theta_2)\, d\theta_1 d\theta_2$$

where

$$h^*(\theta_1,\theta_2) = \begin{cases} h(\pi-\theta_1, \theta_1+\theta_2) & \text{if } 0 \le \theta_1+\theta_2 \le \pi, \\ h(\pi-\theta_1, \theta_1+\theta_2-\pi) & \text{if } \pi < \theta_1+\theta_2 \le 2\pi. \end{cases}$$

Let

$$w_0(\underset{\sim}{t}) = (1+t_1^2)(1+t_2^2)w(\underset{\sim}{t}).$$

LEMMA 3.2. *Suppose $g : \mathbb{R}^2 \longrightarrow \mathbb{R}$ is a continuous function that satisfies*

(i) *gw_0 is bounded on $\mathbb{R}^2$,*

(ii) *$(g \circ f)w_0$ can be extended to a continuous function on $\mathbb{R}^2$, where f is defined in (2.10) and (2.12). Then*

$$\int_{\mathbf{R}^2} g(\underset{\sim}{t})\, w(\underset{\sim}{t})\, d\underset{\sim}{t} = \int_{\mathbf{R}^2} g(f(\underset{\sim}{t}))\, w(\underset{\sim}{t})\, d\underset{\sim}{t}. \tag{3.3}$$

For notational convenience we let

$$(3.4)\qquad <\alpha_1>=\frac{1}{1+t_1^2},\quad <\alpha_2>=\frac{1}{1+t_2^2},\quad <\alpha_1+\alpha_2>=\frac{(t_1+t_2)^2}{(1+t_1^2)(1+t_2^2)}.$$

This notation is related to the notation introduced in Stages 1–3. Using the same change of variables used to derive (1.7) we have

$$(3.5)\quad \text{C.T.}[\alpha_1]^{k_1}[\alpha_2]^{k_2}[\alpha_1+\alpha_2]^{k_3}G = \frac{4^{3a+3b+k_1+k_2+k_3}}{\pi^2}\int_{\mathbf{R}^2} <\alpha_1>^{k_1}<\alpha_2>^{k_2}<\alpha_1+\alpha_2>^{k_3} w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

where $w(\underset{\sim}{t})$ is defined in (2.3) and k_1, k_2, k_3 are nonnegative integers.

LEMMA 3.6.

$$(3.7)\qquad \int_{\mathbf{R}^2}(t_2^2+2t_1t_2-1)<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t}=0,$$

$$(3.8)\qquad \int_{\mathbf{R}^2}\frac{t_1}{t_1+t_2}<\alpha_2> w(\underset{\sim}{t})\,d\underset{\sim}{t}=\frac{5}{4}\int_{\mathbf{R}^2}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t} - \int_{\mathbf{R}^2}<\alpha_1><\alpha_2> w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.9)\qquad \int_{\mathbf{R}^2}\frac{t_1}{t_1-t_2}<\alpha_2> w(\underset{\sim}{t})\,d\underset{\sim}{t}=\frac{3}{4}\int_{\mathbf{R}^2}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.10)\qquad \int_{\mathbf{R}^2}\frac{t_2(t_1+t_2)}{t_2^2+2t_1t_2-1}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t}=\frac{1}{2}\int_{\mathbf{R}^2}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.11)\qquad \int_{\mathbf{R}^2}\frac{1-t_1t_2}{t_1^2+2t_1t_2-1}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t}=-\frac{3}{4}\int_{\mathbf{R}^2}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.12)\qquad \int_{\mathbf{R}^2}\frac{1}{t_1^2+2t_1t_2-1}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t}=-\int_{\mathbf{R}^2}<\alpha_1><\alpha_2> w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.13)\qquad \int_{\mathbf{R}^2}<\alpha_1><\alpha_2><\alpha_1+\alpha_2> w(\underset{\sim}{t})\,d\underset{\sim}{t}=-\frac{3}{4}\int_{\mathbf{R}^2}<\alpha_1>^2 w(\underset{\sim}{t})\,d\underset{\sim}{t} + \frac{3}{2}\int_{\mathbf{R}^2}<\alpha_1><\alpha_2> w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.14)\qquad \int_{\mathbf{R}^2}\frac{1}{t_2^2+2t_1t_2-1}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t}=-\frac{1}{2}\int_{\mathbf{R}^2}<\alpha_1>^2 w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

$$(3.15)\qquad \int_{\mathbf{R}^2}\frac{t_1t_2}{t_2^2+2t_1t_2-1}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t}=\frac{1}{2}\int_{\mathbf{R}^2}<\alpha_1> w(\underset{\sim}{t})\,d\underset{\sim}{t} - \frac{1}{2}\int_{\mathbf{R}^2}<\alpha_1>^2 w(\underset{\sim}{t})\,d\underset{\sim}{t}.$$

Proof. Equation (3.7) follows from (3.3) and

$$(t_2^2 + 2t_1t_2 - 1) <\alpha_1><\alpha_2> \xrightarrow{f} -(t_2^2 + 2t_1t_2 - 1) <\alpha_1><\alpha_2> .$$

We note that $(t_2^2 + 2t_1t_2 - 1) <\alpha_1><\alpha_2>$ satisfies the conditions of Lemma 3.2 since it is equal to $<\alpha_1+\alpha_2> - <\alpha_2>$. Whenever we apply f to a rational function in the remainder of the proof we leave it to the reader to verify that the function involved satisfies the conditions of Lemma 3.2.

Equation (3.8) follows from (3.7) and

$$\frac{t_1}{t_1+t_2} <\alpha_2> + \frac{t_2}{t_1+t_2} <\alpha_1>$$
$$= \frac{1}{2} <\alpha_1><\alpha_2> \{-(t_2^2 + 2t_1t_2 - 1) + 3(1+t_2^2) + 2(1+t_1^2) - 4\}.$$

Similarly (3.9) follows from

$$\frac{t_1}{t_1-t_2} <\alpha_2> + \frac{t_2}{t_2-t_1} <\alpha_1>$$
$$= \frac{1}{2} <\alpha_1><\alpha_2> \{(t_2^2 + 2t_1t_2 - 1) + (1+t_2^2) + 2(1+t_1^2)\}.$$

We have

$$\frac{t_2(t_1+t_2)}{t_2^2 + 2t_1t_2 - 1} <\alpha_1> = <\alpha_1> + \frac{(1-t_1t_2)}{t_2^2 + 2t_1t_2 - 1} <\alpha_1>, \tag{3.16}$$
$$\frac{(1-t_1t_2)}{t_2^2 + 2t_1t_2 - 1} <\alpha_1> \xrightarrow{f} -\frac{t_2(t_1+t_2)}{t_2^2 + 2t_1t_2 - 1} <\alpha_1>$$

and (3.10) follows.

$$\frac{(1-t_1t_2)}{t_1^2 + 2t_1t_2 - 1} <\alpha_1> \xrightarrow{f} \xrightarrow{t_1 \leftrightarrow t_2} -\frac{t_1}{t_1-t_2} <\alpha_2>$$

and (3.11) follows by (3.9).

$$\frac{<\alpha_1>}{t_1^2 + 2t_1t_2 - 1} + <\alpha_1><\alpha_2>$$
$$= \frac{<\alpha_1+\alpha_2>}{t_1^2 + 2t_1t_2 - 1} \xrightarrow{t_1 \leftrightarrow t_2} \xrightarrow{f} \xrightarrow{t_1 \leftrightarrow t_2} -\frac{<\alpha_1+\alpha_2>}{t_1^2 + 2t_1t_2 - 1}$$

and (3.12) follows.

We observe that (3.13) is equivalent to

$$\int_{\mathbf{R}^2} A(\underset{\sim}{t}) w(\underset{\sim}{t})\, d\underset{\sim}{t} = 0,$$

where

$$A(\underset{\sim}{t}) = 4I(\underset{\sim}{t}) + 3 <\alpha_1+\alpha_2>^2 - 6 <\alpha_1><\alpha_1+\alpha_2>,$$
$$I(\underset{\sim}{t}) = <\alpha_1><\alpha_2><\alpha_1+\alpha_2>,$$

since

$$<\alpha_1>^2 \xrightarrow{t_1 \leftrightarrow t_2} \xrightarrow{f} <\alpha_1+\alpha_2>^2,$$
$$<\alpha_1><\alpha_2> \xrightarrow{f} <\alpha_1><\alpha_1+\alpha_2> .$$

$$A(\underset{\sim}{t}) = I(\underset{\sim}{t})(-2 + 3t_1^2 + 6t_1t_2 - 3t_2^2)$$

so that

$$\int_{\mathbf{R}^2} A(\underset{\sim}{t})w(\underset{\sim}{t})\,d\underset{\sim}{t} = -2\int_{\mathbf{R}^2} I(\underset{\sim}{t})(1-3t_1t_2)w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

since $I(\underset{\sim}{t})$ is invariant under $t_1 \leftrightarrow t_2$.

$$(1-t_1t_2) \xrightarrow{f} \frac{t_2(1+t_1^2)}{(t_1+t_2)},$$
$$\frac{t_2(1+t_1^2)}{(t_1+t_2)} + \frac{t_1(1+t_2^2)}{(t_1+t_2)} = (1+t_1t_2).$$

It follows that

$$2\int_{\mathbf{R}^2} I(\underset{\sim}{t})(1-t_1t_2)w(\underset{\sim}{t})\,d\underset{\sim}{t} = \int_{\mathbf{R}^2} I(\underset{\sim}{t})(1+t_1t_2)w(\underset{\sim}{t})\,d\underset{\sim}{t},$$

since $I(\underset{\sim}{t})$ is invariant under f and $t_1 \leftrightarrow t_2$. Therefore,

$$\int_{\mathbf{R}^2} I(\underset{\sim}{t})(1-3t_1t_2)w(\underset{\sim}{t})\,d\underset{\sim}{t} = 0,$$

as required. This completes the proof of (3.13).

Equation (3.14) follows from

$$\frac{<\alpha_1>}{t_2^2+2t_1t_2-1} \xrightarrow{f} -\frac{<\alpha_1>}{t_2^2+2t_1t_2-1} - <\alpha_1>^2 .$$

Finally, (3.15) follows easily from (3.10), (3.14), and (3.16). □

4. Proof of Stages 2–3 and (1.2)–(1.5). In this section we use Lemma 3.6 to prove Stages 2–3, thus completing the proof of (1.2)–(1.4). Finally, we show how (1.5) follows from (1.3) and (1.4). For $k_1, k_2 \geq 0$ we have

$$\begin{aligned}
(4.1)\qquad 0 &= \int_{\mathbf{R}^2} \frac{\partial}{\partial t_1} t_1 <\alpha_1>^{k_1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t} \\
&= -(4a+8b+2k_1+1)\int_{\mathbf{R}^2} <\alpha_1>^{k_1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t} \\
&\quad + 2(2a+4b+k_1+1)\int_{\mathbf{R}^2} <\alpha_1>^{k_1+1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t} \\
&\quad + 2a\int_{\mathbf{R}^2} \frac{t_1}{t_1+t_2} <\alpha_1>^{k_1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t} \\
&\quad + 2b\int_{\mathbf{R}^2} \frac{t_1}{t_1-t_2} <\alpha_1>^{k_1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t} \\
&\quad + 4b\int_{\mathbf{R}^2} \frac{t_1(t_1+t_2)}{t_1^2+2t_1t_2-1} <\alpha_1>^{k_1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t} \\
&\quad + 4b\int_{\mathbf{R}^2} \frac{t_1t_2}{t_2^2+2t_1t_2-1} <\alpha_1>^{k_1}<\alpha_2>^{k_2} w(\underset{\sim}{t})\,d\underset{\sim}{t}.
\end{aligned}$$

Letting $k_1 = 0$, $k_2 = 1$ in (4.1) and using (3.8)–(3.12), we find that

$$\int_{\mathbb{R}^2} <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} = \frac{(3a+3b+2)}{4(a+2b+1)} \int_{\mathbb{R}^2} <\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t}. \tag{4.2}$$

Hence, by (3.5) we have

$$\begin{aligned}
\text{C.T.}\,[\alpha_1][\alpha_2]G &= \frac{4^{3a+3b}}{\pi^2} \int_{\mathbb{R}^2} 16 <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&= \frac{(3a+3b+2)}{(a+2b+1)} \frac{4^{3a+3b}}{\pi^2} \int_{\mathbb{R}^2} 4 <\alpha_1> w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&= \frac{(3a+3b+2)}{(a+2b+1)} \text{C.T.}\,[\alpha_1]G \\
&= \frac{2(3a+3b+1)(3a+3b+2)}{(2a+3b+1)(a+2b+1)} g'(a,b) \qquad \text{(by Stage 1).}
\end{aligned}$$

This completes the proof of Stage 2.

We cannot prove Stage 3 directly. Instead we use (4.1) to get $\int_{\mathbb{R}^2} <\alpha_1>^2 w(\underset{\sim}{t})\, d\underset{\sim}{t}$ in terms of $\int_{\mathbb{R}^2} <\alpha_1> w(\underset{\sim}{t})\, d\underset{\sim}{t}$ and $\int_{\mathbb{R}^2} <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t}$. Then we show how Stage 3 will follow from Stages 1 and 2 using (3.13). Letting $k_1 = 1$, $k_2 = 0$ in (4.1) and using (3.8), (3.9), (3.11), and (3.15) we find that

$$\begin{aligned}
&2(2a+3b+2) \int_{\mathbb{R}^2} <\alpha_1>^2 w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&\quad = \frac{3}{2}(3a+3b+2) \int_{\mathbb{R}^2} <\alpha_1> w(\underset{\sim}{t})\, d\underset{\sim}{t} - 2a \int_{\mathbb{R}^2} <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&\quad = 2(2a+6b+3) \int_{\mathbb{R}^2} <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} \qquad \text{(by (4.2)).}
\end{aligned} \tag{4.3}$$

We have

$$\begin{aligned}
&\text{C.T.}\,[\alpha_1][\alpha_2][\alpha_1+\alpha_2]G \\
&= \frac{4^{3a+3b}}{\pi^2} \int_{\mathbb{R}^2} 64 <\alpha_1><\alpha_2><\alpha_1+\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&= -3 \cdot \frac{4^{3a+3b}}{\pi^2} \int_{\mathbb{R}^2} 16 <\alpha_1>^2 w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&\qquad + 6 \cdot \frac{4^{3a+3b}}{\pi^2} \int_{\mathbb{R}^2} 16 <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} \qquad \text{(by (3.13))} \\
&= \frac{(-3(2a+6b+3)+6(2a+3b+2))}{(2a+3b+2)} \frac{4^{3a+3b}}{\pi^2} \int_{\mathbb{R}^2} 16 <\alpha_1><\alpha_2> w(\underset{\sim}{t})\, d\underset{\sim}{t} \\
&\qquad \text{(by (4.3))} \\
&= \frac{3(2a+1)}{(2a+3b+2)} \text{C.T.}\,[\alpha_1][\alpha_2]G
\end{aligned}$$

$$= \frac{6(3a+3b+2)(3a+3b+1)(2a+1)}{(2a+3b+2)(2a+3b+1)(a+2b+1)} g'(a,b) \qquad \text{(by Stage 2).}$$

This completes the proof of Stage 3.

At the beginning of §2 we showed how Stage 3 implies (1.2). We remark that (1.3) and (1.4) follow from Stages 1 and 2 together with (1.2). We have been unable to find a proof of (1.5) or (1.6) in terms of integrals. However (1.5) follows easily from (1.3) and (1.4). In order to show this we need to recall how the Weyl group acts on polynomials. For $\alpha = k_1\alpha_1 + k_2\alpha_2$, where $k_1, k_2 \in \mathbb{Z}$ and α_1, α_2 are the roots from the root system G_2 as in Fig. 1, we let

$$x^\alpha = x_1^{k_1} x_2^{k_2}.$$

The elements w of the Weyl group W act on monomials by

$$w(x^\alpha) = x^{w(\alpha)},$$

and by linearity on Laurent polynomials that are linear combinations of the x^α. For w in the Weyl group W we have

$$\text{C.T.}\, x^\alpha G = \text{C.T.}\, x^{w(\alpha)} G, \tag{4.4}$$

since G is symmetric with respect to the Weyl group and w does not change the constant term. Utilizing (4.4) we find that the left-hand sides of (1.3)–(1.5) can be written as

$$\text{C.T.}\,[\alpha_1]G = 2\text{C.T.}(1-x_1)G, \tag{4.5}$$

$$\text{C.T.}\,[\alpha_1][\alpha_2]G = 2\text{C.T.}\left(2 - 3x_1 + \frac{x_1}{x_2}\right) G, \tag{4.6}$$

$$\text{C.T.}\,[2\alpha_1+\alpha_2]G = 2\text{C.T.}\left(1 - \frac{x_1}{x_2}\right) G. \tag{4.7}$$

Hence,

$$\begin{aligned}\text{C.T.}\,[2\alpha_1+\alpha_2]G &= 3\text{C.T.}\,[\alpha_1]G - \text{C.T.}\,[\alpha_1][\alpha_2]G \\ &= \left\{\frac{6(3a+3b+1)}{(2a+3b+1)} - \frac{2(3a+3b+2)(3a+3b+1)}{(2a+3b+1)(a+2b+1)}\right\} g(a,b) \\ &\qquad\qquad \text{(by (1.3), (1.4))} \\ &= \frac{2(3a+3b+1)(3b+1)}{(2a+3b+1)(a+2b+1)} g(a,b),\end{aligned}$$

which is (1.5).

5. Other results. In this section we state other results that are similar to (1.3)–(1.6). These may be proved by extending Zeilberger's [12] proof of the ordinary G_2 case as mentioned in §1.

(5.1)

$$\text{C.T.}\,[\alpha_1][\alpha_1-\alpha_2]G = \frac{4(3a+3b+2)(3a+3b+1)(3b+1)}{(2a+3b+2)(2a+3b+1)(a+2b+1)} g(a,b),$$

(5.2)

$$\begin{aligned}&\text{C.T.}\,[\alpha_1][\alpha_1+\alpha_2][\alpha_1-\alpha_2]G \\ &\qquad = \frac{6(4a+3b+4)(3a+3b+2)(3a+3b+1)(3b+1)}{(2a+3b+3)(2a+3b+2)(2a+3b+1)(a+2b+1)} g(a,b),\end{aligned}$$

(5.3)

$$\text{C.T.}\,[\alpha_1][\alpha_2][\alpha_1-\alpha_2]G = 6\left\{\frac{(3a+3b+2)(3a+3b+1)(3b+1)}{(2a+3b+1)(a+2b+2)(a+2b+1)} - \frac{b(4a+3b+4)(3a+3b+2)(3a+3b+1)(3b+1)}{(2a+3b+3)(2a+3b+2)(2a+3b+1)(a+2b+2)(a+2b+1)}\right\} g(a,b),$$

(5.4)

$$\text{C.T.}\,[\alpha_1][\alpha_1-\alpha_2][\alpha_1+2\alpha_2]G = \frac{6(3a+4b+4)(3a+3b+2)(3a+3b+1)(3b+2)(3b+1)}{(2a+3b+3)(2a+3b+2)(2a+3b+1)(a+2b+2)(a+2b+1)}g(a,b),$$

(5.5)

$$\text{C.T.}\,[\alpha_1][\alpha_1-\alpha_2][2\alpha_1+\alpha_2]G = \frac{6(6a^2+20ab+23a+12b^2+28b+16)}{(2a+3b+4)(2a+3b+3)(2a+3b+2)} \cdot \frac{(3a+3b+2)(3a+3b+1)(3b+2)(3b+1)}{(2a+3b+1)(a+2b+2)(a+2b+1)}g(a,b),$$

(5.6)

$$\text{C.T.}\,[\alpha_1][\alpha_1+\alpha_2][\alpha_1-\alpha_2][\alpha_1+2\alpha_2]G = \frac{18(3a+3b+4)(3a+3b+2)(3a+3b+1)}{(2a+3b+4)(2a+3b+3)(2a+3b+2)} \cdot \frac{(2a+2b+3)(3b+2)(3b+1)}{(2a+3b+1)(a+2b+2)(a+2b+1)}g(a,b).$$

Acknowledgments. I would like to thank many people: Richard Askey for introducing me to the Macdonald conjectures, for his conjectures on G_2, and for help and encouragement; Murad Ozaydin for helpful discussions on root systems and Weyl groups; William Long for getting me started on REDUCE; and Dennis Stanton for his suggestions in writing this paper.

REFERENCES

[1] K. AOMOTO, *Jacobi polynomials associated with Selberg's integral*, SIAM J. Math. Anal. 18 (1987), pp. 545–549.

[2] R. ASKEY, *Aomoto's extension of Selberg's integral*, unpublished.

[3] ———, *Integration and Computers*, preprint, to appear in proceedings of a computer algebra conference edited by G. and D. Chudnovsky.

[4] W. N. BAILEY, *Generalized Hypergeometric Series*, Cambridge University Press, London and New York. (Reprinted: Hafner, New York, 1964)

[5] L. HABSIEGER, *La q-conjecture de Macdonald-Morris pour G_2*, C.R. Acad Sci. 303 (1986), pp. 211–213.

[6] K. W. J. Kadell, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal. 19 (1988), pp. 969–986.

[7] ———, private communication.

[8] I. G. MACDONALD, *Some conjectures for root systems*, SIAM J. Math. Anal. 13 (1982), pp. 988–1007.

[9] M. L. MEHTA, *Random Matrices*, Academic Press, New York.

[10] W. G. MORRIS, *Constant term identities for finite and affine root systems*, Ph.D. thesis, Univ. of Wisconsin-Madison, 1982.

[11] A. SELBERG, *Bermerkninger om et multiplet integral*, Norske Mat. Tidsskr. 26 (1944), pp. 71–78.
[12] D. ZEILBERGER, *A proof of the G_2 case of Macdonald's root system-Dyson conjecture*, SIAM J. Math. Anal. 18 (1987), 880–883.
[13] ———, *A unified approach to Macdonald's root-system conjectures*, SIAM J. Math. Anal., 19 (1988), to appear.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

020

UNE q-INTÉGRALE DE SELBERG ET ASKEY*

LAURENT HABSIEGER†

Résumé. Nous prouvons une conjecture de R. Askey ("Some basic hypergeometric extensions of integrals of Selberg and Andrews," *SIAM J. Math. Anal.*, 11 (1980), pp. 938-951), qui propose une q-généralisation de l'intégrale de Selberg:

$$\int_0^1 \cdots \int_0^1 \prod_{1\leq i<j\leq n} |t_i - t_j|^{2z} \prod_{i=1}^{n} t_i^{x-1}(1-t_i)^{y-1}\, dt_1 \cdots dt_n.$$

Nous en déduisons une conjecture de Morris sur le terme constant de:

$$\prod_{j=1}^{l} (x_0/x_j)_a (qx_j/x_0)_b \prod_{1\leq i<j\leq l} (x_i/x_j)_c (qx_j/x_i)_c,$$

où $(x)_k = (1-x)(1-qx)\cdots(1-q^k x)$. En appendice se trouve la preuve d'une autre conjecture Askey, liée à la q-conjecture de Dyson.

Abstract. We prove a conjecture by R. Askey ("Some basic hypergeometric extensions of integrals of Selberg and Andrews," *SIAM J. Math. Anal.*, 11 (1980), pp. 938-951) on a basic extension of Selberg's integral:

$$\int_0^1 \cdots \int_0^1 \prod_{1\leq i<j\leq n} |t_i - t_j|^{2z} \prod_{i=1}^{n} t_i^{x-1}(1-t_i)^{y-1}\, dt_1 \cdots dt_n.$$

We deduce from this a conjecture due to Morris about the constant term in the expansion of

$$\prod_{j=1}^{l} (x_0/x_j)_a (qx_j/x_0)_b \prod_{1\leq i<j\leq l} (x_i/x_j)_c (qx_j/x_i)_c,$$

where $(x)_k = (1-x)(1-qx)\cdots(1-q^k x)$. In the appendix there can be found a proof of another conjecture by Askey related to the Dyson q-conjecture.

Key words. q-analogues, beta and gamma functions, constant term, q-integral, continuous, Macdonald conjecture

AMS(MOS) subject classifications. 33A15, 33A75, 05A19

1. Introduction.

Commençons par définir les notations que nous utiliserons par la suite. Dans tout l'article, q désignera un nombre réel de l'intervalle ouvert $]0, 1[$. En fait, on peut, la plupart du temps, considérer q comme une indéterminée et se placer dans l'algèbre des séries formelles associées à q; les résultats démontrés restent en général valides. Toutefois cette restriction à $]0, 1[$ permet de donner un sens évident à l'expression "faire tendre q vers 1."

Posons $(a)_\infty = \prod_{n=0}^{\infty}(1-aq^n)$ et pour $n \in \mathbf{C}$, $(a)_n = (a)_\infty/(aq^n)_\infty$. La q-fonction gamma, introduite par Jackson [7], est définie sur $\mathbf{C}\backslash\mathbf{Z}^-$ par

$$\Gamma_q(a) = \frac{(q)_\infty}{(q^a)_\infty}(1-q)^{1-a}.$$

Jackson [7] définit également la q-intégration par

$$\int_0^1 f(t)\, d_q t = (1-q)\sum_{n=0}^{\infty} f(q^n)q^n. \tag{1.1}$$

* Received by the editors June 2, 1986; accepted for publication (in revised form) June 12, 1987.

† Département de mathématique, Université Louis-Pasteur, 7, rue René-Descartes, F-67084 Strasbourg, France.

Les dernières notations à introduire pour comprendre la q-généralisation sont, en prenant pour k un entier naturel:

$$\Delta_k^0(t_1,\cdots,t_n)=\prod_{1\leqq i\neq j\leqq n}(\varepsilon_{ij}t_i/t_j)_k,\qquad \text{où } \varepsilon_{ij}=\begin{cases}1 & \text{si } i<j,\\ q & \text{si } i>j,\end{cases}$$

$$\Delta_k(t_1,\cdots,t_n)=\frac{1}{n!}\sum_{\sigma\in S_n}\Delta_k^0(t_{\sigma 1},\cdots,t_{\sigma n}),$$

$$\Delta_k^1(t_1,\cdots,t_n)=\prod_{1\leqq i<j\leqq n}\prod_{l=0}^{k-1}(t_j-q^l t_i)(t_j-q^{-l}t_i).$$

L'intégrale de Selberg est une intégrale à plusieurs variables généralisant la fonction bêta. Elle vaut:

$$\int_0^1\cdots\int_0^1\prod_{1\leqq i<j\leqq n}(t_i-t_j)^{2k}\prod_{i=1}^n t_i^{x-1}(1-t_i)^{y-1}\,dt_1\cdots dt_n$$
$$=\prod_{j=1}^n\frac{\Gamma(x+(j-1)k)\Gamma(y+(j-1)k)\Gamma(jk+1)}{\Gamma(x+y+(n+j-2)k)\Gamma(k+1)}. \tag{1.2}$$

Askey [3] en a proposé plusieurs q-extensions. Nous allons prouver sa première conjecture, à savoir:

$$\int_0^1\cdots\int_0^1\prod_{1\leqq i<j\leqq n}t_i^{2k}\left(\frac{t_j}{t_i}q^{1-k}\right)_{2k}\prod_{i=1}^n t_i^{x-1}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\,d_qt_1\cdots d_qt_n$$
$$=q^{kx\binom{n}{2}+2k^2\binom{n}{3}}\prod_{j=1}^n\frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1)}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)}. \tag{1.3}$$

De manière équivalente, en notant que

$$t_i^{2k}\left(\frac{t_j}{t_i}q^{1-k}\right)_{2k}=(-1)^k(t_it_j)^kq^{-\binom{k}{2}}\left(\frac{t_i}{t_j}\right)_k\left(\frac{t_j}{t_i}q\right)_k,$$

nous montrerons que:

$$\int_{[0,1]^n}\Delta_k^0(\mathbf{t})\prod_{i=1}^n t_i^{x+(n-1)k-1}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\,d_qt$$
$$=(-1)^{k\binom{n}{2}}q^{\binom{n}{2}\binom{k}{2}+k\binom{n}{2}x+2k^2\binom{n}{3}}$$
$$\cdot\prod_{j=1}^n\frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1)}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)}, \tag{1.4}$$

où x et y sont deux nombres complexes de parties réelles suffisamment grandes pour assurer la convergence de l'intégrale et où k est un entier naturel.

Regardons maintenant ce qui se passe lorsque q tend vers 1. Askey [4] a prouvé que $\lim_{q\to 1}\Gamma_q(x)=\Gamma(x)$. De plus on vérifie sans peine que $\lim_{q\to 1}(qt)_\infty/(q^yt)_\infty=\lim_{q\to 1}\sum_{n=0}^\infty(q^{1-y})_n(tq^y)^n/(q)_n=\sum_{n=0}^\infty(1-y)_nt^n/n!=(1-t)^{y-1}$ et le théorème de convergence des sommes de Riemann nous assure que $\lim_{q\to 1}\int_0^1 f(t)\,d_qt=\int_0^1 f(t)\,dt$. Ces remarques nous permettent de vérifier que lorsqu'on fait tendre q vers 1 dans (1.3) et (1.4), on retrouve bien (1.2).

Dans le deuxième paragraphe, nous étudierons quelques propriétés utiles de la q-fonction gamma et de la q-intégration puis nous calculerons Δ_k. Nous montrerons ensuite au troisième paragraphe qu'il suffit alors d'évaluer la quantité

$$\int_{[0,1]^n}\Delta_k^1(t)\prod_{i=1}^n t_i^{x-1}\frac{(t_iq)_\infty}{(t_iq^y)_\infty}\,d_qt. \tag{1.5}$$

Nous suivrons alors la preuve de Selberg pour montrer que (1.5) peut se mettre sous la forme

$$\prod_{j=1}^{n} \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)}{\Gamma_q(x+y+(n+j-2)k)} \cdot \frac{R(q^x, q^y)}{P(q^y)},$$

où R et P sont des polynômes à deux et une variables respectivement.

A ce stade, dans la preuve classique, on utilisait la symétrie en x et y pour montrer que $R(X, Y)/P(Y)$ était une constante, indépendante de X et Y. Ici cet argument n'est plus valable car (1.5) n'est pas symétrique en x et y. Nous démontrerons toutefois au quatrième paragraphe que l'on a $R(X, Y)/P(Y) = C_n(k)X^{k\binom{n}{2}}$, où $C_n(k)$ est une constante, indépendante de X et Y, grâce à l'utilisation appropriée de développements limités.

Le cinquième paragraphe sera consacré au calcul de $C_n(k)$, suivant deux méthodes. Au sixième paragraphe, nous utiliserons (1.4) pour prouver une conjecture due à Morris [9]:

$$CT \prod_{j=1}^{l} \left(\frac{x_0}{x_j}\right)_a \left(q\frac{x_j}{x_0}\right)_b \prod_{1\leqq i<j\leqq l} \left(\frac{x_i}{x_j}\right)_c \left(q\frac{x_j}{x_i}\right)_c = \prod_{j=0}^{l-1} \frac{(q)_{a+b+jc}(q)_{(j+1)c}}{(q)_{a+jc}(q)_{b+jc}(q)_{jc}}.$$

Enfin, en appendice, nous prouverons une autre conjecture d'Askey, tirée du même article que (1.3), en montrant qu'elle est essentiellement équivalente à la q-conjecture de Dyson.

Après rédaction de cet article, Kadell nous a communiqué son mémoire [8] dans lequel il démontre également la conjecture d'Askey sur la q-intégrale de Selberg. Il reprend, en les généralisant, les méthodes développées par Aomoto [2]. Les techniques de démonstration de Kadell ne recouvrent absolument pas les nôtres.

2. Calculs préliminaires. Tout d'abord, montrons quelques formules relatives à la q-fonction gamma.

PROPRIÉTÉS. On a:

$$\Gamma_q(n+1) = \frac{(q)_n}{(1-q)^n} \quad \text{pour } n \in \mathbf{N}, \tag{2.1}$$

$$\frac{\Gamma_q(a+n)}{\Gamma_q(a)} = \frac{(q^a)_n}{(1-q)^n} \quad \text{et} \quad \frac{\Gamma_q(a-n)}{\Gamma_q(a)} = \frac{(1-q)^n}{(q^{a-n})_n} \quad \text{pour } n \in \mathbf{N}, \tag{2.2}$$

$$\lim_{x\to\infty} \frac{\Gamma_q(x+a)}{\Gamma_q(x)} = (1-q)^{-a}, \tag{2.3}$$

$$\lim_{\varepsilon\to 0} \frac{1-q^{\varepsilon}}{1-q}\Gamma_q(-a+\varepsilon) = \frac{(1-q)^a(-1)^a q^{\binom{a+1}{2}}}{(q)_a} \quad \text{pour } a \in \mathbf{N}. \tag{2.4}$$

Preuves. Propriété (2.1) est une conséquence triviale des définitions de Γ_q et $(q)_n$. Pour $n \in \mathbf{N}$,

$$\frac{\Gamma_q(a+n)}{\Gamma_q(a)} = \frac{(1-q)^{1-a-n}(q)_\infty/(q^{a+n})_\infty}{(1-q)^{1-a}(q)_\infty/(q^a)_\infty}$$

$$= (1-q)^{-n}\frac{(q^a)_\infty}{(a^{a+n})_\infty} = \frac{(q^a)_n}{(1-q)^n}.$$

La deuxième partie de (2.2) est une conséquence de la première.

On a:

$$\frac{\Gamma_q(x+a)}{\Gamma_q(x)} = (1-q)^{-a}\frac{(q^x)_\infty}{(q^{x+a})_\infty}$$

et

$$\lim_{x\to+\infty}(q^x)_\infty = \lim_{x\to+\infty}(q^{x+a})_\infty = (0)_\infty = 1,$$

ce qui entraîne (2.3).

De même,

$$\begin{aligned}\Gamma_q(-a+\varepsilon) &= (1-q)^{1+a-\varepsilon}\frac{(q)_\infty}{(q^{-a+\varepsilon})_\infty}\\ &= \frac{(1-q)^{1+a-\varepsilon}}{1-q^\varepsilon}\cdot\frac{(q)_\infty}{(q^{1+\varepsilon})_\infty}\cdot\frac{1}{\prod_{l=1}^{a}(1-q^{-l+\varepsilon})},\end{aligned}$$

donc

$$\lim_{\varepsilon\to 0}\frac{1-q^\varepsilon}{1-q}\Gamma_q(-a+\varepsilon) = \frac{(1-q)^a}{\prod_{l=1}^{a}(1-q^{-l})} = \frac{(1-q)^a(-1)^a q^{\binom{a+1}{2}}}{(q)_a}.$$

Le point de départ de la théorie des fonctions hypergéométriques basiques est le théorème q-binômial:

$$\frac{(ax)_\infty}{(x)_\infty} = \sum_{n=0}^{\infty}\frac{(a)_n}{(q)_n}x^n. \tag{2.5}$$

En fait il est équivalent à la formule intégrale fondamentale:

$$\int_0^1 t^{x-1}\frac{(qt)_\infty}{(q^y t)_\infty}\,d_q t = \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}. \tag{2.6}$$

Pour le voir, il suffit d'appliquer les définitions:

$$\begin{aligned}\int_0^1 t^{x-1}\frac{(qt)_\infty}{(q^y t)_\infty}\,d_q t &= (1-q)\sum_{n=0}^{\infty} q^{nx}\frac{(q^{n+1})_\infty}{(q^{n+y})_\infty} \quad \text{(d'après (1.1))}\\ &= (1-q)\sum_n \frac{(q)_\infty}{(q^y)_\infty}\frac{(q^y)_n}{(q)_n}(q^x)^n\\ &= (1-q)\frac{(q)_\infty}{(q^y)_\infty}\frac{(q^{x+y})_\infty}{(q^x)_\infty} \quad \text{(grâce à (2.5))}\\ &= \frac{\Gamma_q(x)\Gamma_q(y)}{\Gamma_q(x+y)}, \quad \text{par définition de } \Gamma_q.\end{aligned}$$

On remarquera que la formule (2.6) est une q-généralisation de la formule d'Euler: $\int_0^1 t^{x-1}(1-t)^{y-1}\,dt = \Gamma(x)\Gamma(y)/\Gamma(x+y)$.

Notons aussi l'égalité suivante, valable pour $\sigma \in S_n$:

$$\int_{[0,1]^n} f(\sigma\mathbf{t})\,d_q\mathbf{t} = \int_{[0,1]^n} f(t)\,d_q t. \tag{2.7}$$

En effet $\int_{[0,1]^n} f(\sigma\mathbf{t})\,d_q\mathbf{t} = (1-q)^n\sum_{m\in\mathbf{N}^n}\prod_i q^{m_{\sigma i}}\times f(q^{m_{\sigma 1}},\cdots,q^{m_{\sigma n}}) = (1-q)^n\sum_{\mu\in\mathbf{N}^n}\cdot\prod_i q^{\mu_i}\times f(q^{\mu_1},\cdots,q^{\mu_n}) = \int_{[0,1]^n} f(\mathbf{t})\,d_q\mathbf{t}$. Passons maintenant au calcul de Δ_k.

PROPOSITION. *On a la formule explicite suivante*:

$$\Delta_k(\mathbf{t}) = \frac{\Gamma_{q^k}(n+1)}{n!} \prod_{1\le i\ne j\le n} \left(\frac{t_i}{t_j}\right)_k . \tag{2.8}$$

Démonstration. Posons

$$R(\mathbf{t}) = \frac{n!\Delta_k(\mathbf{t})}{\prod_{1\le i\ne j\le n}(t_i/t_j)_k}.$$

On a

$$R(t) = \sum_{\sigma\in S_n} \frac{\Delta_k^0(\sigma\mathbf{t})}{\prod_{i\ne j}(t_i/t_j)_k} = \sum_\sigma \prod_{i<j} \frac{(qt_{\sigma j}/t_{\sigma i})_k}{(t_{\sigma j}/t_{\sigma i})_k}$$
$$= \sum_\sigma \prod_{i<j} \frac{1-q^k t_{\sigma j}/t_{\sigma i}}{1-t_{\sigma j}/t_{\sigma i}} = \sum_\sigma \varepsilon(\sigma) \prod_{i<j} \frac{q^k t_{\sigma j} - t_{\sigma i}}{t_j - t_i},$$

en notant $\varepsilon(\sigma)$ la signature de la permutation σ. Posons $V(\mathbf{t}) = \prod_{1\le i<j\le n}(t_j - t_i)$. On voit que le produit

$$R(\mathbf{t})V(\mathbf{t}) = \sum_\sigma \varepsilon(\sigma) \prod_{i<j} (q^k t_{\sigma j} - t_{\sigma i})$$

est un polynôme antisymétrique, qui est de degré au plus $\binom{n}{2}$. C'est donc un multiple scalaire de $V(\mathbf{t})$, ou encore $R(\mathbf{t})$ est une constante en $\mathbf{t}$. Pour évaluer cette constante, on calcule le coefficient de $t_n^{n-1}\cdots t_2^1 t_1^0$ dans $RV(\mathbf{t})$. Notons $I(\sigma)$ le nombre d'inversions de la permutation σ, de sorte que $\varepsilon(\sigma) = (-1)^{I(\sigma)}$. On vérifie alors aisément que la contribution du terme $\varepsilon(\sigma)\prod_{i<j}(q^k t_{\sigma j} - t_{\sigma i})$ au coefficient de $t_n^{n-1}\cdots t_2^1 t_1^0$ dans la somme $RV(\mathbf{t})$ est donnée par $\varepsilon(\sigma)q^{k\binom{n}{2}-kI(\sigma)}(-1)^{I(\sigma)}$, c'est-à-dire $q^{k\binom{n}{2}-kI(\sigma)}$. De là, $R(\mathbf{t})q^{-k\binom{n}{2}}$ n'est autre que la fonction génératrice, sur le groupe des permutations, du nombre des inversions. La variable utilisée étant q^{-k}, cette fonction génératrice est égale à $\Gamma_{q^{-k}}(n+1)$, comme il est bien connu (cf. [6]), ce qui donne $R(\mathbf{t}) = \Gamma_{q^k}(n+1)$.

3. Factorisation de la q-intégrale de Selberg. Prenons comme point de départ la forme (1.4) de la conjecture. Puisque pour tout $\sigma\in S_n$,

$$\int_{[0,1]^n} \Delta_k^0(\sigma\mathbf{t}) \prod_i t_i^{x+(n-1)k-1} \frac{(t_i q)}{(t_i q^y)_\infty} d_q\mathbf{t}$$
$$= \int_{[0,1]^n} \Delta_k^0(\sigma\mathbf{t}) \prod_i t_{\sigma i}^{x+(n-1)k-1} \frac{(t_{\sigma i} q)_\infty}{(t_{\sigma i} q^y)_\infty} d_q\mathbf{t}$$
$$= \int_{[0,1]^n} \Delta_k^0(\mathbf{t}) \prod_i t_i^{x+(n-1)k-1} \frac{(t_i q)_\infty}{(t_i q^y)_\infty} d_q\mathbf{t} \quad \text{grâce à (2.7)},$$

on a forcément

$$\int_{[0,1]^n} \Delta_k^0(\mathbf{t}) \prod_i t_i^{x+(n-1)k-1} \frac{(t_i q)_\infty}{(t_i q^y)_\infty} d_q\mathbf{t} = \int_{[0,1]^n} \Delta_k(\mathbf{t}) \prod_i t_i^{x+(n-1)k-1} \frac{(t_i q)_\infty}{(t_i q^y)_\infty} d_q\mathbf{t}.$$

Or

$$\Delta_k(t) = \frac{\Gamma_{q^k}(n+1)}{n!} \prod_{i\ne j} \left(\frac{t_i}{t_j}\right)_k \quad \text{(grâce à (2.8))}$$
$$= \frac{\Gamma_{q^k}(n+1)}{n!} \prod_{i<j} \prod_{l=0}^{k-1} \left(1 - q^l \frac{t_i}{t_j}\right)\left(1 - q^l \frac{t_j}{t_i}\right)$$
$$= \frac{\Gamma_{q^k}(n+1)}{n!} \prod_{i<j} \prod_{l=0}^{k-1} \left(-\frac{q^l}{t_i t_j}\right)(t_i - q^l t_j)(t_i - q^{-l} t_j),$$

c'est-à-dire:

$$\Delta_k(\mathbf{t}) = \frac{\Gamma_{q^k}(n+1)}{n!}(-1)^{k\binom{n}{2}} q^{\binom{k}{2}\binom{n}{2}} \Delta_k^1(\mathbf{t}) \prod_{i=1}^{n} t_i^{-(n-1)k}. \tag{3.1}$$

Posons donc:

$$F_n(x, y, k) = q^{-kx\binom{n}{2}} \int_{[0,1]^n} \Delta_k^1(\mathbf{t}) \prod_i t_i^{x-1} \frac{(t_i q)_\infty}{(t_i q^y)_\infty} d_q\mathbf{t}. \tag{3.2}$$

En confrontant ce qui précède à (1.4), on voit qu'on est ramené à prouver le théorème suivant.

THÉORÈME. *La formule suivante est valide.*

$$F_n(x, y, k) = \frac{n! q^{2k^2\binom{n}{3}}}{\Gamma_{q^k}(n+1)} \prod_{j=1}^{n} \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)\Gamma_q(jk+1)}{\Gamma_q(x+y+(n+j-2)k)\Gamma_q(k+1)}. \tag{3.3}$$

Démonstration. On suit tout d'abord la preuve classique: on développe $\Delta_k^1(\mathbf{t}) = \sum_{\alpha \in \mathbf{N}^n} c(\alpha) t^\alpha$, où les $c(\alpha)$ sont presque tous nuls, de sorte que

$$\begin{aligned} F_n(x, y, k) &= q^{-kx\binom{n}{2}} \sum_\alpha c(\alpha) \int_{[0,1]^n} \prod_i t_i^{x+\alpha_i-1} \frac{(qt_i)_\infty}{(q^y t_i)_\infty} d_q\mathbf{t} \\ &= q^{-kx\binom{n}{2}} \sum_\alpha c(\alpha) \prod_i \frac{\Gamma_q(x+\alpha_i)\Gamma_q(y)}{\Gamma_q(x+y+\alpha_i)}, \quad \text{grâce à (2.6).} \end{aligned}$$

Or $\Delta_k(\mathbf{t})$ est symétrique, par construction, donc $\Delta_k^1(\mathbf{t})$ est aussi symétrique, grâce à (3.1). On a donc

$$F_n(x, y, k) = q^{-kx\binom{n}{2}} \sum_{0 \leqq \beta_1 \leqq \cdots \leqq \beta_n} c'(\beta) \prod_i \frac{\Gamma_q(x+\beta_i)\Gamma_q(y)}{\Gamma_q(x+y+\beta_i)}, \tag{3.4}$$

où

$$\begin{aligned} c'(\beta) &= c(\beta) \cdot \#\{\alpha \in \mathbf{N}^n : \exists \sigma \in S_n \,|\, \sigma\alpha = \beta\} \\ &\quad (= c(\beta) \cdot n! \text{ si } \beta \text{ a toutes ses coordonnées distinctes}). \end{aligned}$$

Or $\Delta_k^1(\mathbf{t})$ est homogène de degré $2k\binom{n}{2}$, donc $kn(n-1) = \beta_1 + \cdots + \beta_n \leqq n\beta_n$ et ainsi $\beta_n \geqq k(n-1)$.

De plus, $\Delta_k^1(t_1, \cdots, t_j)$ divise $\Delta_k^1(t_1, \cdots, t_n)$, pour $1 \leqq j \leqq n$, et ainsi la puissance de t_j dans $\Delta_k^1(t_1, \cdots, t_n)$ est au moins celle de t_j dans $\Delta_k^1(t_1, \cdots, t_j)$. On en déduit que:

$$\text{pour } j \in \{1, \cdots, n\}, \qquad \beta_j \geqq k(j-1). \tag{3.5}$$

En outre,

$$\begin{aligned} \Delta_k^1\left(\frac{1}{t_1}, \cdots, \frac{1}{t_n}\right) &= \prod_{i<j} \prod_l \left(\frac{1}{t_i} - \frac{q^l}{t_j}\right)\left(\frac{1}{t_i} - \frac{q^{-l}}{t_j}\right) \\ &= \prod_{i<j} \prod_l \left(-\frac{q^l}{t_i t_j}\right)\left(-\frac{q^{-l}}{t_i t_j}\right)(t_i - q^{-l}t_j)(t_i - q^l t_j) \\ &= \prod_i t_i^{-2(n-1)k} \Delta_k^1(\mathbf{t}). \end{aligned}$$

Alors $2(n-1)k - \beta_{n+1-j} \geqq (j-1)k$, en utilisant (3.5). Donc

$$\beta_j \leqq 2(n-1)k - (n+1-j-1)k = (n+j-2)k \quad \text{pour } 1 \leqq j \leqq n. \tag{3.6}$$

De (3.5) et (2.2), on déduit que

$$\frac{\Gamma_q(x+\beta_j)}{\Gamma_q(x+(j-1)k)}$$

est un polynôme en q^x, pour $1 \leqq j \leqq n$. Il résulte alors de (3.6) et (2.2) que, pour $1 \leqq j \leqq n$,

$$\frac{\Gamma_q(x+y+(n+j-2)k)}{\Gamma_q(x+y+\beta_j)}$$

est un polynôme en q^{x+y}, de degré $(n+j-2)k-\beta_j$. En reportant dans (3.4), on trouve

$$F_n(x, y, k) = q^{-k\binom{n}{2}x} \sum_{\beta} c'(\beta) \prod_j \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y)}{\Gamma_q(x+y+(n+j-2)k)} R_\beta(q^x, q^y),$$

où $R_\beta(X, Y)$ est un polynôme en deux variables, de degré en Y égal à

$$\sum_j ((n+j-2)k-\beta_j) = k\binom{n}{2}.$$

Le produit mis en évidence ne dépend plus de β; donc $F_n(x, y, k)$ peut se mettre sous la forme

$$\prod_j \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y)}{\Gamma_q(x+y+(n+j-2)k)} \cdot \frac{R_0(q^x, q^y)}{q^{k\binom{n}{2}x}}.$$

De plus,

$$\frac{\Gamma_q(y)}{\Gamma_q(y+(j-1)k)} = \frac{(1-q)^{(j-1)k}}{(q^y)_{(j-1)k}} \quad \text{pour } 1 \leqq j \leqq n.$$

Donc

$$F_n(x, y, k) = \prod_j \frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)}{\Gamma_q(x+y+(n+j-2)k)} \cdot \frac{R(q^x, q^y)}{q^{k\binom{n}{2}x} P(q^y)}, \tag{3.7}$$

où R est un polynôme en deux variables, de degré en Y inférieur ou égal à $k\binom{n}{2}$ et où P est un polynôme en Y de degré $k\binom{n}{2}$:

$$P(Y) = \prod_{j=1}^{n} (Y)_{(j-1)k} = \prod_{j=1}^{n-1} (Y)_{jk}. \tag{3.8}$$

Il faut maintenant démontrer que $R(X, Y)/(X^{k\binom{n}{2}}P(Y))$ est une constante, indépendante de X et Y.

4. Simplification de $R(X, Y)/(X^{k\binom{n}{2}}P(Y))$. Prenons $F_n(x, y, k)$ sous la forme (1.1):

$$F_n(x, y, k) = (1-q)^n q^{-k\binom{n}{2}x} \sum_{m \in \mathbf{N}^n} \Delta_k^1(q^{m_1}, \cdots, q^{m_n}) \prod_{i=1}^{n} q^{m_i x} \frac{(q^{m_i+1})_\infty}{(q^{m_i+y})_\infty}.$$

La fonction Δ_k^1 est symétrique. Elle est de plus nulle si deux des variables sont égales (en fait pour $k \geqq 1$, le cas $k=0$ étant trivial). Donc

$$\begin{aligned} F_n(x, y, k) = n!\, q^{-k\binom{n}{2}x}(1-q)^n \sum_{0 \leqq m_1 \leqq \cdots \leqq m_n} \Delta_k^1(q^{m_1}, \cdots, q^{m_n}) \\ \cdot \prod_i q^{m_i x} \frac{(q^{m_i+1})_\infty}{(q^{m_i+y})_\infty}. \end{aligned} \tag{4.1}$$

Or

$$\Delta_k^1(q^{m_1}, \cdots, q^{m_n}) = \prod_{i<j} \prod_l (q^{m_i} - q^{l+m_j})(q^{m_i} - q^{-l+m_j});$$

donc s'il existe une paire $\{i,j\}\subset\{1,\cdots,n\}$ telle que $|m_i-m_j|<k$, alors $\Delta_k^1(q^{m_1},\cdots,q^{m_n})=0$. Ainsi, en supposant les m_i ordonnés, on a

(4.2) $$\Delta_k^1(q^{m_1},\cdots,q^{m_n})\neq 0\Leftrightarrow 0\leqq m_1\leqq m_2-k\leqq\cdots\leqq m_n-(n-1)k.$$

(4.3) En particulier (4.2) entraîne que $m_1+\cdots+m_n\geqq k\binom{n}{2}$. Il y a de plus égalité si et seulement si $m_i=(i-1)k$ pour $1\leqq i\leqq n$.

Faisons tendre x vers $+\infty$: grâce à (4.3),

$$F_n(x,y,k)\to n!(1-q)^n\Delta_k^1(1,q^k,\cdots,q^{(n-1)k})\prod_i\frac{(q^{(i-1)k+1})_\infty}{(q^{(i-1)k+y})_\infty}.$$

En utilisant (2.3), on voit que

$$\prod_j\frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)}{\Gamma_q(x+y+(n+j-2)k)}\cdot\frac{R(q^x,q^y)}{q^{k\binom{n}{2}x}P(q^y)}\sim C_1(y)\frac{R(q^x,q^y)}{q^{k\binom{n}{2}x}}.$$

On en déduit que nécessairement $X^{k\binom{n}{2}}$ divise $R(X,Y)$. On peut donc écrire $F_n(x,y,k)$ sous la forme

$$\prod_{j=1}^n\frac{\Gamma_q(x+(j-1)k)\Gamma_q(y+(j-1)k)}{\Gamma_q(x+y+(n+j-2)k)}\cdot\frac{Q(q^x,q^y)}{P(q^y)},$$

où $P(Y)=\prod_{j=1}^{n-1}(Y)_{jk}$.

Pour $a\in\{0,\cdots,n-2\}$ et $b\in\{0,\cdots,k-1\}$, posons $Y_{a,b}=q^{-(ak+b)}$.

(4.4) On remarque que les $Y_{a,b}$ sont les seuls zéros de P et que la multiplicité de $Y_{a,b}$ dans P est $n-1-a$.

Prenons $y=-ak-b+\varepsilon$ où $\varepsilon\in]0,\frac{1}{2}[$, $a\in\{0,\cdots,n-2\}$, $b\in\{0,\cdots,k-1\}$ et x tel que $x-[x]\in]0,\frac{1}{2}[$. La seconde condition ne fait que nous assurer qu'il n'y a aucun problème de définition et ne nuit pas à la généralité du problème. On fait tendre ε vers zéro. D'après (2.4),

(4.5) $$\prod_j\Gamma_q((j-1)k-ak-b+\varepsilon)\sim C_2(a,b)(1-q^\varepsilon)^{-a-1},$$

car $(j-1)k-ak-b\leqq 0\Leftrightarrow j\leqq a+1$. Grâce à (4.4), on a:

(4.6) $$P(q^{-ak-b+\varepsilon})\sim C_3(a,b)(1-q^\varepsilon)^{n-1-a}.$$

Dans (4.1), on pose $m_j=p_j+(j-1)k$, ce qui est licite, grâce à (4.2). On a:

$$\prod_j(q^{(j-1)k-ak-b+\varepsilon+p_j})_\infty\sim C_4(a,b,p)(1-q^\varepsilon)^{\omega(p)}\quad\text{avec }\omega(p)\leqq a+1,$$

car $\{j:(j-1)k-ak-b+p_j\leqq 0\}\subset\{j:(j-1)k-ak-b\leqq 0\}=\{1,\cdots,a+1\}$.

Récrivant (4.1) à l'aide des p_j, on trouve (pour $k\geqq 1$):

(4.7)
$$F_n(x,y,k)=n!(1-q)^n\sum_{0\leqq p_1\leqq\cdots\leqq p_n}\Delta_k^1(q^{p_1},\cdots,q^{p_n+(n-1)k})\cdot\prod_i q^{p_ix}\frac{(q^{p_i+(i-1)k+1})_\infty}{(q^{p_i+(i-1)k+y})_\infty}.$$

Dans cette somme, seul le dénominateur varie quand y tend vers $-ak-b$ et donc, en utilisant ce qui précède,

(4.8) $$F_n(x,-ak-b+\varepsilon,k)\sim C_5(a,b,x)(1-q^\varepsilon)^{-\omega}\quad\text{avec }\omega\leqq a+1.$$

En combinant (4.5), (4.6) et (4.8), on obtient

$$Q(q^x,q^{-ak-b+\varepsilon})\sim C_6(a,b,x)(1-q^\varepsilon)^{n-\omega}\quad\text{avec }\omega\leqq a+1.$$

Ceci montre que $Y_{a,b}$ est zéro de $Q(X, Y)$, d'ordre au moins $n-(a+1)$. D'après (4.4), on peut en déduire que $P(Y)$ divise $Q(X, Y)$. Or, d'après (3.8), $d_Y^0 R(X, Y) \leqq k\binom{n}{2} = d^0 P$. Ceci prouve que $Q(X, Y)/P(Y)$ est un polynôme en X, disons $Q_0(X)$.

Pour déterminer Q_0, prenons $y = -(n-1)k+\varepsilon$ et faisons tendre ε vers zéro. On a alors:

$$\prod_j \frac{\Gamma_q(x+(j-1)k)\Gamma_q(\varepsilon-(n-j)k)}{\Gamma_q(x+(j-1)k+\varepsilon)} \sim C_7(1-q^\varepsilon)^{-n}. \tag{4.9}$$

De plus, pour $0 \leqq p_1 \leqq \cdots \leqq p_n$ et $p_n > 0$, on a:

$$\prod_j (q^{p_j-(n-j)k+\varepsilon})_\infty^{-1} = o((1-q^\varepsilon)^{-n})$$

et

$$\prod_j (q^{-(n-j)k+\varepsilon})_\infty^{-1} \sim C_8(1-q^\varepsilon)^{-n}.$$

Donc, grâce à (4.7),

$$F_n(x, -(n-1)k+\varepsilon, k) \sim C_9(1-q^\varepsilon)^{-n}, \tag{4.10}$$

où C_9 est indépendante de q^x car le terme prépondérant de la somme est celui correspondant à $p_1 = \cdots = p_n = 0$. En confrontant (4.9) et (4.10), on trouve donc $Q_0(X) = C_9/C_7$: c'est bien une constante indépendante de x. Posons désormais $Q_0(X) = C_n(k)$.

5. Calcul de la constante $C_n(k)$. Il existe plusieurs manières pour déterminer $C_n(k)$. Suivons tout d'abord la preuve classique. Pour $\lambda \in \mathbf{C}$,

$$\int_0^1 t^{x-1}t^\lambda \, d_q t = \frac{1-q}{1-q^{x+\lambda}} \qquad (\mathrm{Re}\,(x+\lambda) > 0), \tag{5.1}$$

donc

$$\lim_{x\to 0} \frac{1-q^x}{1-q} \int_0^1 t^{x-1}t^\lambda \, d_q t = 0 \quad \text{si } \mathrm{Re}\,\lambda > 0.$$

Ainsi on a le q-analogue de $\lim_{x\to 0} x\int_0^1 t^{x-1}f(t)\,dt = f(0)$, à savoir:

$$\lim_{x\to 0} \frac{1-q^x}{1-q} \int_0^1 t^{x-1}f(t) \, d_q t = f(0) \quad \text{si } f \text{ est continue en 0.}$$

Par symétrie, on a la formule correspondant à (4.1):

$$F_n(x, y, k) = q^{-k\binom{n}{2}x} n! \int_0^1 \int_{t_n}^1 \cdots \int_{t_2}^1 \Delta_k^1(t_1, \cdots, t_n) \prod_{i=1}^n t_i^{x-1} \frac{(qt_i)_\infty}{(q^y t_i)_\infty} \, d_q t,$$

d'où

$$\lim_{x\to 0} \frac{1-q^x}{1-q} F_n(x, y, k)$$
$$= n! \int_0^1 \int_{t_{n-1}}^1 \cdots \int_{t_2}^1 \Delta_k^1(t_1, \cdots, t_{n-1}) \prod_{i=1}^{n-1} t_i^{2k-1} \frac{(qt_i)_\infty}{(q^y t_i)_\infty} \, d_q t,$$

car

$$\prod_{i=1}^{n-1} \prod_{l=0}^{k-1} (0-q^l t_i)(0-q^{-l}t_i) = \prod_{i=1}^{n=1} t_i^{2k}.$$

Ainsi

$$\lim_{x\to 0}\frac{1-q^x}{1-q}F_n(x,y,k)=nF_{n-1}(2k,y,k)q^{2k^2\binom{n-1}{2}}.$$

On a alors

$$C_n(k)\cdot\prod_{j=1}^{n-1}\frac{\Gamma_q(jk)\Gamma_q(y+jk)}{\Gamma_q(y+(n+j-1)k)}\cdot\frac{\Gamma_q(y)}{\Gamma_q(y+(n-1)k)}$$
$$=q^{2k^2\binom{n-1}{2}}nC_{n-1}(k)\prod_{j=1}^{n-1}\frac{\Gamma_q((j+1)k)\Gamma_q(y+(j-1)k)}{\Gamma_q(y+(n+j-1)k)};$$

d'où

$$C_n(k)=n!\frac{\Gamma_q(nk)}{\Gamma_q(k)}q^{2k^2\binom{n-1}{2}}C_{n-1}(k)=\cdots=n!\prod_{j=1}^{n}\frac{\Gamma_q(jk)}{\Gamma_q(k)}q^{2k^2\binom{n}{3}},$$

car $C_1(k)=1$. Or

$$\prod_j\frac{\Gamma_q(jk+1)}{\Gamma_q(k+1)}=\prod_j\frac{\Gamma_q(jk)}{\Gamma_q(k)}\cdot\prod_j\frac{1-q^{jk}}{1-q^k}=\prod_j\frac{\Gamma_q(jk)}{\Gamma_q(k)}\cdot\Gamma_{q^k}(n+1),$$

donc

$$C_n(k)=\frac{n!}{\Gamma_{q^k}(n+1)}\prod_{j=1}^{n}\frac{\Gamma_q(jk+1)}{\Gamma_q(k+1)},$$

ce qui achève la démonstration du théorème (3.3).

On peut également trouver $C_n(k)$ de manière plus explicite, en précisant les identités (4.9) et (4.10). En effet, pour $y=-(n-1)k+\varepsilon$ et ε tendant vers zéro, on a, à partir de (4.1):

$$F_n(x,-(n-1)k+\varepsilon,k)$$
$$\sim n!\Delta_k^1(1,\cdots,q^{(n-1)k})\prod_{j=1}^{n}\frac{(q^{(j-1)k+1})_\infty}{(q^{-(n-j)k+\varepsilon})_\infty}(1-q)^n$$
$$\sim n!\Delta_k^1(1,\cdots,q^{(n-1)k})\prod_j\frac{\Gamma_q(-(n-j)k+\varepsilon)(1-q)^{-(n-1)k+\varepsilon}}{\Gamma_q((j-1)k+1)}$$
$$\sim\frac{n!\Delta_k^1(1,\cdots,q^{(n-1)k})(1-q)^{-2k\binom{n}{2}}}{\prod_j\Gamma_q((j-1)k+1)}\prod_j\Gamma_q(-(n-j)k+\varepsilon).$$

D'autre part, on sait que

$$F_n(x,y,k)\sim\prod_{j=1}^{n}\frac{\Gamma_q(x+(j-1)k)\Gamma_q(-(n-j)k+\varepsilon)}{\Gamma_q(x+(j-1)k+\varepsilon)}C_n(k).$$

On en déduit que:

$$C_n(k)=\frac{n!\Delta_k^1(1,\cdots,q^{(n-1)k})(1-q)^{-2k\binom{n}{2}}}{\prod_j\Gamma_q((j-1)k+1)}. \tag{5.2}$$

Rappelons que

$$\Delta_k^1(1,\cdots,q^{(n-1)k})=\prod_{0\leqq i<j\leqq n-1}\prod_{l=0}^{k-1}(q^{ik}-q^{jk+l})(q^{ik}-q^{jk-l}).$$

On pourra sortir de ce produit une puissance de q dont l'exposant vaut:

$$\sum_{0\leqq i<j\leqq n-1}\sum_{l=0}^{k-1} 2ik = \sum_{0\leqq i<j\leqq n-1} 2ik^2 = 2k^2\sum_{i=0}^{n-1} i(n-1-i)$$

$$= k^2 n(n-1)\left(n-1-\frac{2n-1}{3}\right) = 2k^2\binom{n}{3}.$$

On en déduit que:

$$\Delta_k^1(1,\cdots,q^{(n-1)k}) = q^{2k^2\binom{n}{3}} \prod_{0\leqq i<j\leqq n-1}\prod_{l=0}^{k-1}(1-q^{(j-i)k+l})(1-q^{(j-i)k-l})$$

$$= q^{2k^2\binom{n}{3}} \prod_{l=0}^{k-1}\prod_{r=1}^{n-1}(1-q^{rk+l})^{n-r}\prod_{l=1}^{k}\prod_{r=0}^{n-2}(1-q^{rk+l})^{n-r-1}$$

$$= q^{2k^2\binom{n}{3}} \prod_{j=1}^{n}\frac{(q)_{jk-1}}{(q)_{k-1}}\prod_{j=1}^{n-1}(q)_{jk}$$

$$= q^{2k^2\binom{n}{3}} \prod_{j=1}^{n-1}\frac{\Gamma_q(jk)\Gamma_q((j-1)k+1)}{\Gamma_q(k)}(1-q)^{2(j-1)k}.$$

En remplaçant dans (5.2), ceci nous donne:

$$C_n(k) = n!\,q^{2k^2\binom{n}{3}}\prod_j \frac{\Gamma_q(jk)}{\Gamma_q(k)} = \frac{n!\,q^{2k^2\binom{n}{3}}}{\Gamma_{q^k}(n+1)}\prod_j\frac{\Gamma_q(jk+1)}{\Gamma_q(k+1)},$$

pour le mêmes raisons que précédemment.

Le lecteur attentif notera que l'on peut aussi obtenir (5.2) en faisant tendre x vers $+\infty$.

6. Application à une conjecture de Morris. Il s'agit de calculer le terme constant de

$$\prod_{i=1}^{l}\left(\frac{t_0}{t_i}\right)_a\left(q\frac{t_i}{t_0}\right)_b\prod_{1\leqq i\neq j\leqq l}\left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_c.$$

On pose $x_i = q^{-a}t_i/t_0$, ce qui ne change pas le terme constant. Alors

$$\left(\frac{t_0}{t_i}\right)_a\left(q\frac{t_i}{t_0}\right)_b = \prod_{l=0}^{a-1}\left(1-\frac{q^{l-a}}{x_i}\right)\prod_{l=1}^{b}(1-q^{l+a}x_i)$$

$$= (-1)^a q^{-\binom{a+1}{2}}x_i^{-a}(qx_i)_{a+b}.$$

On a donc

$$CT\prod_{i=1}^{l}\left(\frac{t_0}{t_i}\right)_a\left(q\frac{t_i}{t_0}\right)_b\prod_{i\neq j}\left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_c$$

$$= (-1)^{la}q^{-l\binom{a+1}{2}}CT\prod_{i\neq j}\left(\varepsilon_{ij}\frac{x_i}{x_j}\right)_c\prod_i x_i^{-a}\frac{(qx_i)_\infty}{(q^{a+b+1}x_i)_\infty}.$$

Donnons maintenant une généralisation de (5.1). Comme me l'a fait remarquer Askey dans une correspondance privée, celle-ci correspond au q-analogue de la notion d'intégrale généralisée introduite par Hadamard. On pose, pour $\lambda\in\mathbf{C}$:

$$\int_0^1 t^{\lambda-1}\,d_qt = \frac{1-q}{1-q^\lambda}.$$

Ainsi, si f est une série formelle, on peut définir $\int_0^1 t^{x-1} f(t)\, d_q t$, pour $x \in \mathbf{C}\backslash\mathbf{Z}$. On aura de plus:

$$\lim_{x\to 0} \frac{1-q^x}{1-q} \int_0^1 t^{x-1} f(t)\, d_q t = CTf, \tag{6.1}$$

où le symbole $\int$ désigne l'opérateur formel défini ci-dessus. On choisit pour paramètres dans la q-intégrale de Selberg: $n = l$, $x = -(n-1)k - a + \varepsilon$, $\varepsilon \in]0, 1[$, $y = a + b + 1$, $k = c$. On fait tendre ε vers zéro. D'après (6.1),

$$\begin{aligned}
&CT \prod_{i\neq j} \left(\varepsilon_{ij} \frac{x_i}{x_j}\right)_c \prod_i x_i^{-a} \frac{(qx_i)_\infty}{(q^{a+b+1}x_i)_\infty} \\
&= \lim_{\varepsilon\to 0} \left(\frac{1-q^\varepsilon}{1-q}\right)^n \int_{[0,1]^n} \Delta_c^0(\mathbf{t}) \prod_i t_i^{-a+\varepsilon-1} \frac{(qt_i)_\infty}{(q^{a+b+1}t_i)_\infty}\, d_q\mathbf{t} \\
&= (-1)^{c\binom{n}{2}} q^{\binom{n}{2}\binom{c}{2} + c(-(n-1)c-a)\binom{n}{2} + 2c^2\binom{n}{3}} \lim_{\varepsilon\to 0} \left(\frac{1-q^\varepsilon}{1-q}\right)^n \\
&\quad \cdot \prod_{j=1}^n \frac{\Gamma_q(-a-(n-j)c+\varepsilon)\Gamma_q(a+b+(j-1)c+1)\Gamma_q(jc+1)}{\Gamma_q(b+(j-1)c+1+\varepsilon)\Gamma_q(c+1)},
\end{aligned}$$

d'après (1.4). En effet, le membre de gauche de (1.4) n'était défini que pour Re x assez grand. Avec la nouvelle définition introduite ci-dessus, on peut, par analyticité, l'étendre à tout $\mathbf{C}\backslash\mathbf{Z}$. Or

$$\lim_{\varepsilon\to 0} \frac{1-q^\varepsilon}{1-q} \Gamma_q(-a-jc+\varepsilon) = \frac{(1-q)^{a+jc}(-1)^{a+jc} q^{\binom{a+jc+1}{2}}}{(q)_{a+jc}},$$

grâce à (2.4), et on a le système d'égalités:

$$\begin{aligned}
\Gamma_q(a+b+jc+1) &= (1-q)^{-a-b-jc}(q)_{a+b+jc}, \\
\Gamma_q(b+jc+1) &= (1-q)^{-b-jc}(q)_{b+jc}, \\
\Gamma_q((j+1)c+1) &= (1-q)^{-(j+1)c}(q)_{(j+1)c}, \\
\Gamma_q(c+1) &= (1-q)^{-c}(q)_c.
\end{aligned}$$

On trouve donc:

$$\begin{aligned}
&CT \prod_{i=1}^n \left(\frac{t_0}{t_i}\right)_a \left(q\frac{t_i}{t_0}\right)_b \prod_{1\leqq i\neq j\leqq n} \left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_c \\
&= (-1)^{na+c\binom{n}{2}+na+c\binom{n}{2}} q^\alpha \prod_{j=0}^{n-1} \frac{(q)_{a+b+jc}(q)_{(j+1)c}}{(q)_{a+jc}(q)_{b+jc}(q)_c},
\end{aligned}$$

où

$$\begin{aligned}
\alpha &= -n\binom{a+1}{2} + \binom{n}{2}\binom{c}{2} - c(a+(n-1)c)\binom{n}{2} + 2c^2\binom{n}{3} \\
&\quad + n\binom{a+1}{2} + c\left(a+\frac{1}{2}\right)\binom{n}{2} + \frac{c^2}{2}\frac{n(n-1)(2n-1)}{6} \\
&= \binom{n}{2}\binom{c}{2} - (n-1)c^2\binom{n}{2} + 2c^2\binom{n}{3} + \frac{c}{2}\binom{n}{2} + \frac{c^2 n(n-1)(2n-1)}{12} \\
&= \frac{c}{2}\binom{n}{2}\left(c-1-2c(n-1) + \frac{4c(n-2)}{3} + 1 + \frac{c(2n-1)}{3}\right) \\
&= \frac{c}{2}\binom{n}{2}\left(c(3-2n) + \frac{c}{3}(6n-9)\right) = 0.
\end{aligned}$$

On trouve donc bien la forme "Cauchy-Selberg" de la conjecture B de Morris [9]:

$$CT \prod_{i=1}^{l} \left(\frac{t_0}{t_i}\right)_a \left(q\frac{t_i}{t_0}\right)_b \prod_{1\leqq i\neq j\leqq l} \left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_c = \prod_{j=0}^{l-1} \frac{(q)_{a+b+jc}(q)_{(j+1)c}}{(q)_{a+jc}(q)_{b+jc}(q)_c}.$$

On remarquera que lorsque $a=b=0$, ceci fournit la q-conjecture de Dyson avec tous les paramètres égaux (cf. [1]).

Appendice: Une autre conjecture d'Askey. Posons

$$F_n(a_1,\cdots,a_n;x)=\int_{[0,1]^n} \prod_{i\neq j}\left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_{a_i} \prod_i t_i^{x-1}\frac{(t_iq)_\infty}{(t_iq^{a_i+1-x})_\infty} d_q\mathbf{t},$$

$$\text{où } \varepsilon_{ij}=\begin{cases}1 & \text{si } i<j,\\ q & \text{si } i>j.\end{cases}$$

On définit aussi:

$$G_n(a_1,\cdots,a_n;x)=\frac{F_n(a_1,\cdots,a_n;x)}{F_n(0,\cdots,0;x)}$$

et

$$H_n(a_1,\cdots,a_n;x)=\frac{\Gamma_q(a_1+\cdots+a_n+1-x)}{\Gamma_q(a_1+1)\cdots\Gamma_q(a_n+1)\Gamma_q(1-x)}.$$

THÉORÈME. *Pour* $(a_1,\cdots,a_n)\in\mathbf{N}^n$ *et* $x\in\mathbf{C}\backslash\mathbf{N}^*$, *on a la formule suivante*:

$$G_n(a_1,\cdots,a_n;x)=H_n(a_1,\cdots,a_n;x).$$

La conjecture initiale d'Askey [3, Conjecture 4] disait que:

$$F_n(a_1,\cdots,a_n;x)=H_n(a_1,\cdots,a_n;x)\cdot[\Gamma_q(x)\Gamma_q(1-x)]^n.$$

Il y a équivalence avec le théorème car

$$F_n(0,\cdots,0;x)=\int_{[0,1]^n}\prod_i t_i^{x-1}\frac{(t_iq)_\infty}{(t_iq^{1-x})_\infty}d_q\mathbf{t}=\prod_{i=1}^n\Gamma_q(x)\Gamma_q(1-x),$$

d'après (2.6). Askey avait mis en évidence le fait que sa conjecture entrainait la q-conjecture de Dyson à n paramètres:

$$CT\prod_{1\leqq i\neq j\leqq n}\left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_{a_i}=\frac{(q)_{a_1+\cdots+a_n}}{(q)_{a_1}\cdots(q)_{a_n}}.$$

Nous allons en fait montrer que le théorème est une conséquence de la q-conjecture de Dyson à $n+1$ paramètres. Puisque cette conjecture a été prouvée (cf. Bressoud et Zeilberger, [5]), ceci démontrera le théorème. Il va sans dire qu'une preuve directe du théorème serait bien plus intéressante mais nous n'en avons pas trouvé.

D'après (2.2), $H_n(a_1,\cdots,a_n;x)$ est une fraction rationnelle en q^x, définie sur $\mathbf{C}\backslash\mathbf{N}^*$. Développons

$$\prod_{i\neq j}\left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_{a_i}=\sum_\alpha A(\alpha)\mathbf{t}^\alpha,$$

où la somme est prise sur les n-uplets d'entiers relatifs $\alpha=(\alpha_1,\cdots,\alpha_n)$ tels que $\alpha_1+\cdots+\alpha_n=0$. Alors:

$$\begin{aligned}
&G_n(a_1,\cdots,a_n;x)\\
&=\sum_\alpha A(\alpha)\int_{[0,1]^n}\prod_i t_i^{x+\alpha_i-1}\frac{(t_iq)_\infty}{(t_iq^{a_i+1-x})_\infty}\,d_q\mathbf{t}\cdot[\Gamma_q(x)\Gamma_q(1-x)]^{-n}\\
&=\sum_\alpha A(\alpha)\prod_i\frac{\Gamma_q(x+\alpha_i)}{\Gamma_q(x)}\,\frac{\Gamma_q(a_i+1-x)}{\Gamma_q(1-x)}\,\frac{1}{\Gamma_q(a_i+1+\alpha_i)}.
\end{aligned}$$

Cette dernière relation montre bien que G_n est définie sur $\mathbf{C}\backslash\mathbf{Z}$. C'est de plus une fraction rationnelle en q^x. Montrons que G_n admet un prolongement continu aux entiers négatifs, qui coïncide avec H_n. Ceci suffira à prouver que G_n et H_n sont égales car deux fractions rationnelles qui coïncident en une infinité de valeurs distinctes sont forcément égales.

Soient $a_0\in\mathbf{N}$, $x\in]0,1[$. On a:

$$G_n(a_1,\cdots,a_n;-a_0+x)=\sum_\alpha A(\alpha)\prod_i\frac{\Gamma_q(\alpha_i-a_0+x)\Gamma_q(a_i+1+a_0-x)}{\Gamma_q(-a_0+x)\Gamma_q(1+a_0-x)\Gamma_q(a_i+\alpha_i+1)}.$$

Si $\alpha\geqq 0$, alors

$$\frac{\Gamma_q(\alpha-a_0+x)}{\Gamma_q(-a_0+x)}=\frac{(q^{-a_0+x})_\alpha}{(1-q)^\alpha}\to\frac{(q^{-a_0})_\alpha}{(1-q)^\alpha}$$

quand x tend vers zéro et

$$\frac{(q^{-a_0})_\alpha}{(1-q)^\alpha}=\begin{cases}0 & \text{si } \alpha\geqq a_0,\\ \dfrac{(-1)^\alpha q^{\binom{\alpha}{2}-a_0\alpha}}{(1-q)^\alpha}\cdot\dfrac{(q)_{a_0}}{(q)_{a_0-\alpha}} & \text{si } \alpha\leqq a_0.\end{cases}$$

Si $\alpha\leqq 0$, alors

$$\frac{\Gamma_q(\alpha-a_0-x)}{\Gamma_q(-a_0+x)}=\frac{(1-q)^{-\alpha}}{(q^{\alpha-a_0+x})_{-\alpha}}\to\frac{(1-q)^{-\alpha}}{(q^{\alpha-a_0})_{-\alpha}},$$

quand x tend vers zéro. Et de même:

$$\frac{(1-q)^{-\alpha}}{(q^{\alpha-a_0})_{-\alpha}}=\frac{(-1)^\alpha q^{-\binom{\alpha+1}{2}-(a_0-\alpha)\alpha}}{(1-q)^\alpha}\cdot\frac{(q)_{a_0}}{(q)_{a_0-\alpha}}.$$

Ainsi, dans tous les cas, on trouve que:

$$\lim_{x\to 0}\frac{\Gamma_q(\alpha-a_0+x)}{\Gamma_q(-a_0+x)}=\frac{(-1)^\alpha q^{\binom{\alpha}{2}-a_0\alpha}}{(1-q)^\alpha}\cdot\frac{(q)_{a_0}}{(q)_{a_0-\alpha}}.$$

On en déduit que $G_n(a_1,\cdots,a_n;-a_0+x)$ admet pour limite quand x tend vers zéro:

$$\begin{aligned}
&G_n(a_1,\cdots,a_n;-a_0)\\
&=\sum_\alpha A(\alpha)\prod_i\frac{(-1)^{\alpha_i}q^{\binom{\alpha_i}{2}-a_0\alpha_i}}{(1-q)^{\alpha_i}}\,\frac{(q)_{a_0}}{(q)_{a_0-\alpha_i}}\,\frac{\Gamma_q(a_i+1+a_0)}{\Gamma_q(a_i+1+\alpha_i)\Gamma_q(a_0+1)}\\
&=\sum_\alpha A(\alpha)\prod_i(-1)^{\alpha_i}q^{\binom{\alpha_i}{2}}\frac{(q)_{a_0+a_i}}{(q)_{a_0-\alpha_i}(q)_{a_i+\alpha_i}},
\end{aligned}$$

grâce à (2.1) et au fait que $\prod q^{-a_0\alpha_i}=q^{-a_0\sum\alpha_i}=1$.

Or, pour $i \in \{1, \cdots, n\}$, on a la formule suivante, conséquence classique de (2.5):

$$\left(\frac{t_0}{t_i}\right)_{a_0}\left(q\frac{t_i}{t_0}\right)_{a_i} = \sum_{\alpha_i} \frac{(-1)^{\alpha_i} q^{\binom{\alpha_i}{2}} (q)_{a_0+a_i}}{(q)_{a_0-\alpha_i}(q)_{a_i+\alpha_i}} \left(\frac{t_0}{t_i}\right)^{\alpha_i}.$$

On voit donc que:

$$\begin{aligned} G_n(a_1, \cdots, a_n; -a_0) &= CT \prod_{0 \leqq i \neq j \leqq n} \left(\varepsilon_{ij}\frac{t_i}{t_j}\right)_{a_i} \\ &= \frac{(q)_{a_0+\cdots+a_n}}{(q)_{a_0}\cdots(q)_{a_n}} \\ &= \frac{\Gamma_q(a_0+\cdots+a_n+1)}{\Gamma_q(a_0+1)\cdots\Gamma_q(a_n+1)} \quad \text{grâce à (2.1).} \end{aligned}$$

Ainsi $G_n(a_1, \cdots, a_n; -a_0) = H_n(a_1, \cdots, a_n; -a_0)$ et le théorème est prouvé.

Remerciements. J'aimerais remercier M. Foata pour les nombreuses lectures qu'il m'a communiquées et l'oreille attentive qu'il m'a tendue. J'aimerais aussi remercier M. Bressoud pour le cours intéressant qu'il nous a dispensé cette année à Strasbourg, en particulier pour son remarquable exposé sur l'intégrale de Selberg.

BIBLIOGRAPHIE

[1] G. E. ANDREWS, *Notes on the Dyson conjecture*, SIAM J. Math. Anal., 11 (1980), pp. 787-792.
[2] AOMOTO, *Jacobi polynomials associated with Selberg integrals*, SIAM J. Math. Anal., 18 (1987), pp. 545-549.
[3] R. ASKEY, *Some basic hypergeometric extensions of integrals of Selberg and Andrews*, SIAM J. Math. Anal., 11 (1980), pp. 938-951.
[4] ———, *The q-gamma and q-beta functions*, Appl. Anal., 8 (1978), pp. 125-141.
[5] D. BRESSOUD AND D. ZEILBERGER, *Proof of Andrews's q-Dyson conjecture*, Discrete Math., 54 (1985), pp. 201-224.
[6] L. COMTET, *Analyse Combinatoire*, Vol. 2, Paris, Presses Universitaires de France, 1970, p. 80.
[7] F. H. JACKSON, *On q-definite integrals*, Quart. J. Pure Appl. Math., 41 (1910), pp. 193-203.
[8] K. KADELL, *A proof of Askey's conjectured q-analogue of Selberg's integral and a conjecture of Morris*, SIAM J. Math. Anal., 19 (1988), pp. 969-986.
[9] W. MORRIS, *Constant term identities for finite and affine root systems, conjectures and theorems*, Ph.D. thesis, University of Wisconsin, Madison, Wisconsin, 1982.
[10] A. SELBERG, *Bemerkninger on et multipelt integral*, Norsk. Mat. Tidsskr., 26 (1944), pp. 71-78.

SIAM J. MATH. ANAL.
Vol. 19, No. 6, November 1988

021

A PROOF OF RAMANUJAN'S IDENTITY BY USE OF LOOP INTEGRALS*

KATSUHISA MIMACHI†

Abstract. Ramanujan's identity means the following:

$$\sum_{n=-\infty}^{+\infty} \frac{(a;q)_n}{(b;q)_n} x^n = \frac{(ax;q)_\infty (q/ax;q)_\infty (b/a;q)_\infty (q;q)_\infty}{(x;q)_\infty (b/ax;q)_\infty (q/a;q)_\infty (b;q)_\infty},$$

where $(a;q)_\infty = \prod_{j=0}^{+\infty}(1-aq^j)$, $(a;q)_n = (a;q)_\infty/(aq^n;q)_\infty$ for $-\infty < n < +\infty$, and $|b/a| < |x| < 1$, $|q| < 1$. This identity plays an important role in the theory of "q-analysis" (see, for example, [1], [3]). Various proofs of it are known ([2], [4], etc.). The aim of this paper is to derive the identity by another method, that of loop integrals.

Key words. Ramanujan's ${}_1\Psi_1$ identity, residue calculus

AMS(MOS) subject classifications. primary 33A30; secondary 10A45

1. Notation.

Set

$$C_n := \{\rho_n \exp(\sqrt{-1}\varphi) \mid \rho_n := \tfrac{1}{2}(|q|^n + |q|^{n+1}),\ 0 \leqq \varphi \leqq 2\pi\},$$

$$\tilde{C}_n := \{\tilde{\rho}_n \exp(\sqrt{-1}\varphi) \mid \tilde{\rho}_n := \tfrac{1}{2}|a/b|(|q|^{-n-1} + |q|^{-n}),\ 0 \leqq \varphi \leqq 2\pi\}$$

in the usual counterclockwise direction. Define

$$f(t) := \frac{(tq^2/a;q)_\infty (a/tq;q)_\infty}{(tx;q)_1 (1/t;q)_\infty (tb/a;q)_\infty},$$

$$F(t) := \frac{(a;q)_\infty (b/a;q)_\infty (q;q)_\infty (tq^2/a;q)_\infty (a/tq;q)_\infty}{(b;q)_\infty (a/q;q)_\infty (q^2/a;q)_\infty (tx;q)_1 (1/t;q)_\infty (tb/a;q)_\infty}$$

$$= \frac{(a;q)_\infty (b/a;q)_\infty (q;q)_\infty}{(b;q)_\infty (a/q;q)_\infty (q^2/a;q)_\infty} f(t),$$

$$I(C) := \frac{1}{2\pi\sqrt{-1}} \int_C f(t)\, dt,$$

$$\operatorname*{Res}_{t=y} \varphi(t) := \text{"the residue of } \varphi(t) \text{ at } t = y\text{"}.$$

2. The function $F(t)$ has simple poles at $t = q^j$, $t = a/bq^j$ $(j = 0, 1, 2, \cdots)$, and $t = 1/x$. The infinite point ∞ and the origin 0 are essential singularities.

LEMMA 1. *If* $|b/a| < |x| < 1$, $|a| < 1$, $|q| < |b|$, *then we have*

$$\sum_{n=-\infty}^{+\infty} \frac{(a;q)_n}{(b;q)_n} x^n - \frac{(ax;q)_\infty (q/ax;q)_\infty (b/a;q)_\infty (q;q)_\infty}{(x;q)_\infty (b/ax;q)_\infty (q/a;q)_\infty (b;q)_\infty}$$

$$= \sum_{j=0}^{+\infty} \operatorname*{Res}_{t=q^j} F(t) + \operatorname*{Res}_{t=1/x} F(t) + \sum_{j=0}^{+\infty} \operatorname*{Res}_{t=a/bq^j} F(t).$$

* Received by the editors April 22, 1987; accepted for publication (in revised form) December 15, 1987.
† Department of Mathematics, Faculty of Science, Nagoya University, Furocho, Chikusa-Ku, Nagoya 464, Japan.

Proof. By summing up the part of nonnegative powers in the left-hand side, we have

$$\begin{aligned}\sum_{n=0}^{+\infty}\frac{(a;q)_n}{(b;q)_n}x^n &= \frac{(a;q)_\infty}{(b;q)_\infty}\sum_{n=0}^{+\infty}\sum_{j=0}^{+\infty}\frac{(b/a;q)_j(aq^n)^j}{(q;q)_j}x^n\\ &= \frac{(a;q)_\infty}{(b;q)_\infty}\sum_{j=0}^{+\infty}\sum_{n=0}^{+\infty}(xq^j)^n\frac{a_j\cdot(b/a;q)_j}{(q;q)_j}\\ &= \frac{(a;q)_\infty}{(b;q)_\infty}\sum_{j=0}^{+\infty}\frac{a^j\cdot(b/a;q)_j}{(xq^j;q)_1(q;q)_j}\\ &= \sum_{j=0}^{+\infty}\operatorname*{Res}_{t=q^j}F(t).\end{aligned}$$

The above expansions are valid when $|a|<1$, $|x|<1$. Similarly,

$$\begin{aligned}\sum_{-\infty}^{-1}\frac{(a;q)_n}{(b;q)_n}x^n &= \sum_{n=1}^{+\infty}\frac{(q/b;q)_n}{(q/a;q)_n}\left(\frac{b}{ax}\right)^n\\ &= \frac{(q/b;q)_\infty}{(q/a;q)_\infty}\sum_{n=1}^{+\infty}\frac{(q^{n+1}/a;q)_\infty}{(q^{n+1}/b;q)_\infty}\left(\frac{b}{ax}\right)^n\\ &= \frac{(q/b;q)_\infty}{(q/a;q)_\infty}\sum_{n=1}^{+\infty}\sum_{j=0}^{+\infty}\frac{(b/a;q)_j(q^{n+1}/b)^j}{(q;q)_j}\left(\frac{b}{ax}\right)^n\\ &= \frac{(q/b;q)_\infty}{(q/a;q)_\infty}\sum_{j=0}^{+\infty}\sum_{n=1}^{+\infty}\left(\frac{bq^j}{ax}\right)^n\frac{(b/a;q)_j(q/b)^j}{(q;q)_j}\\ &= \frac{(q/b;q)_\infty}{(q/a;q)_\infty}\left(\frac{b}{ax}\right)\sum_{j=0}^{+\infty}\frac{(b/a;q)_j(q^2/b)^j}{(bq^j/ax;q)_1(q;q)_j}\\ &= \sum_{j=0}^{+\infty}\operatorname*{Res}_{t=a/bq^j}F(t).\end{aligned}$$

We note that the above expansions are valid when $|q/b|<1$, $|b/ax|<1$. On the other hand, we have simply that

$$\operatorname*{Res}_{t=1/x}F(t) = -\frac{(ax;q)_\infty(q/ax;q)_\infty(b/a;q)_\infty(q;q)_\infty}{(x;q)_\infty(b/ax;q)_\infty(q/a;q)_\infty(b;q)_\infty},$$

which completes the proof. □

What remains to be done is to estimate the effect of essential singularities at ∞ and zero. We obtain the following lemma.

LEMMA 2. *Under the condition* $|a|<1$, $|q^2|<|b|$, $|b/a|<|x|<1$, *we have*

$$\sum_{j=0}^{+\infty}\operatorname*{Res}_{t=q^j}F(t)+\operatorname*{Res}_{t=1/x}F(t)+\sum_{j=0}^{+\infty}\operatorname*{Res}_{t=a/bq^j}F(t)=0.$$

Proof. Due to the definition, we only have to prove

$$\sum_{j=0}^{+\infty}\operatorname*{Res}_{t=q^j}f(t)+\operatorname*{Res}_{t=1/x}f(t)+\sum_{j=0}^{+\infty}\operatorname*{Res}_{t=a/bq^j}f(t)=0.$$

Cauchy's theorem shows

$$\sum_{j=0}^{m}\operatorname*{Res}_{t=q^j}f(t)+\operatorname*{Res}_{t=1/x}f(t)+\sum_{j=0}^{n}\operatorname*{Res}_{t=a/bq^j}f(t)=I(\tilde{C}_n)-I(C_m).$$

Therefore the proof is completed from the following lemma.

LEMMA 3. *Under the condition* $|b/a|<|x|<1$,

(1) *if* $|a|<1$, *then* $|I(C_m)|\to 0$ *for* $m\to+\infty$;
(2) *if* $|q^2|<|b|$, *then* $|I(\tilde{C}_n)|\to 0$ *for* $n\to+\infty$.

Proof. (1) For $m=1,2,3,\cdots$, we have from the definition

$$f(|q|^m t)=(a/q)^m(tq^2|q|^m/aq^m;q)_m(aq^m/tq|q|^m;q)_\infty(tq^2|q|^m/a;q)_\infty \cdot(tq|q|^m/q^m;q)_m^{-1}(q^m/t|q|^m;q)_\infty^{-1}(tb|q|^m/a;q)_\infty^{-1}(tx|q|^m;q)_1^{-1}.$$

Hence for $0\leqq\varphi\leqq 2\pi$,

$$\begin{aligned}|f(\rho_m e^{\sqrt{-1}\varphi})| &= |f(\rho_0|q|^m e^{\sqrt{-1}\varphi})| \\ &= |a/q|^m\times|(\rho_0 q^2|q|^m e^{\sqrt{-1}\varphi}/aq^m;q)_m| \\ &\quad\times|(aq^m/\rho_0 q|q|^m e^{\sqrt{-1}\varphi};q)_\infty(\rho_0 q^2|q|^m e^{\sqrt{-1}\varphi}/a;q)_\infty| \\ &\quad\times|(\rho_0 q|q|^m e^{\sqrt{-1}\varphi}/q^m;q)_m(q^m/\rho_0|q|^m e^{\sqrt{-1}\varphi};q)_\infty|^{-1} \\ &\quad\times|(\rho_0 b|q|^m e^{\sqrt{-1}\varphi}/a;q)_\infty(\rho_0 x|q|^m e^{\sqrt{-1}\varphi};q)_1|^{-1}.\end{aligned} \tag{1}$$

For each factor in the right-hand side, we have the following estimates:

$$\begin{aligned}|(\rho_0 q^2|q|^m e^{\sqrt{-1}\varphi}/aq^m;q)_m| &= \prod_{j=0}^{m-1}\left|1-\frac{\rho_0 q^{2+j}|q|^m e^{\sqrt{-1}\varphi}}{aq^m}\right| \\ &\leqq \prod_{j=0}^{m-1}\left(1+\left|\frac{\rho_0 q^{2+j}}{a}\right|\right)\leqq\prod_{j=0}^{+\infty}\left(1+\left|\frac{\rho_0 q^{2+j}}{a}\right|\right),\end{aligned} \tag{2}$$

$$|(aq^m/\rho_0 q|q|^m e^{\sqrt{-1}\varphi};q)_\infty|\leqq\prod_{j=0}^{+\infty}\left|1-\frac{aq^{m+j}}{\rho_0 q|q|^m e^{\sqrt{-1}\varphi}}\right|\leqq\prod_{j=0}^{+\infty}\left(1+\left|\frac{aq^{j-1}}{\rho_0}\right|\right), \tag{3}$$

$$\begin{aligned}|(\rho_0 q^2|q|^m e^{\sqrt{-1}\varphi}/a;q)_\infty| &\leqq \prod_{j=0}^{+\infty}\left|1-\frac{\rho_0 q^{2+j}|q|^m e^{\sqrt{-1}\varphi}}{a}\right| \\ &\leqq \prod_{j=0}^{+\infty}\left(1+\left|\frac{\rho_0 q^{m+2+j}}{a}\right|\right)\leqq\prod_{j=0}^{+\infty}\left(1+\left|\frac{\rho_0 q^{2+j}}{a}\right|\right),\end{aligned} \tag{4}$$

$$\begin{aligned}|(\rho_0 q|q|^m e^{\sqrt{-1}\varphi}/q^m;q)_m| &= \prod_{j=0}^{m-1}\left|1-\frac{\rho_0 q^{1+j}|q|^m e^{\sqrt{-1}\varphi}}{q^m}\right| \\ &\geqq \prod_{j=0}^{m-1}(1-|\rho_0 q^{1+j}|)\geqq\prod_{j=0}^{+\infty}(1-|\rho_0 q^{1+j}|)>0,\end{aligned} \tag{5}$$

$$\begin{aligned}|(q^m/\rho_0|q|^m e^{\sqrt{-1}\varphi};q)_\infty| &= \prod_{j=0}^{+\infty}\left|1-\frac{q^{m+j}}{\rho_0|q|^m e^{\sqrt{-1}\varphi}}\right|\geqq\left|1-\frac{q^m}{\rho_0|q|^m e^{\sqrt{-1}\varphi}}\right|\prod_{j=1}^{+\infty}\left(1-\left|\frac{q^j}{\rho_0}\right|\right) \\ &\geqq\frac{1-\rho_0}{\rho_0}\prod_{j=1}^{+\infty}\left(1-\left|\frac{q^j}{\rho_0}\right|\right)>0,\end{aligned} \tag{6}$$

$$|(\rho_0 x|q|^m e^{\sqrt{-1}\varphi};q)_1|\geqq 1-|\rho_0 xq^m|\geqq 1-|\rho_0 x|>0, \tag{7}$$

$$\begin{aligned}|(\rho_0 b|q|^m e^{\sqrt{-1}\varphi}/a;q)_\infty| &= \prod_{j=0}^{+\infty}\left|1-\frac{\rho_0 bq^j|q|^m e^{\sqrt{-1}\varphi}}{a}\right| \\ &\geqq\prod_{j=0}^{+\infty}\left(1-\left|\frac{\rho_0 bq^{m+j}}{a}\right|\right)\geqq\prod_{j=0}^{+\infty}\left(1-\left|\frac{\rho_0 bq^j}{a}\right|\right)>0.\end{aligned} \tag{8}$$

By (1)-(8), there exists a positive number M such that

$$|f(\rho_m e^{\sqrt{-1}\varphi})|\leqq M\cdot\left|\frac{a}{q}\right|^m \qquad (0\leqq\varphi\leqq 2\pi). \tag{9}$$

Hence,

$$|I(C_m)| = \left|\frac{1}{2\pi\sqrt{-1}}\int_{C_m} f(t)\,dt\right| \leqq \frac{\rho_m}{2\pi}\int_0^{2\pi} |f(\rho_m e^{\sqrt{-1}\varphi})| \cdot |d\varphi|$$

$$\leqq \frac{\rho_m}{2\pi} \cdot \operatorname*{Max}_{0\leqq\varphi\leqq 2\pi} |f(\rho_m e^{\sqrt{-1}\varphi})| \cdot 2\pi \leqq \rho_m \cdot M \cdot \left|\frac{a}{q}\right|^m \leqq \rho_0 \cdot M \cdot |a|^m.$$

Consequently, for $|a|<1$, $|I(C_m)|\to 0$ if $m\to+\infty$.

(2) The proof is similar to (1). □

THEOREM. *Under the condition* $|b/a|<|x|<1$, $|q|<1$, *we have*

$$\sum_{n=-\infty}^{+\infty} \frac{(a;q)_n}{(b;q)_n} x^n = \frac{(ax;q)_\infty (q/ax;q)_\infty (b/a;q)_\infty (q;q)_\infty}{(x;q)_\infty (b/ax;q)_\infty (q/a;q)_\infty (b;q)_\infty}. \tag{10}$$

Proof. Lemmas 1 and 2 verify (10), if $|b/a|<|x|<1$, $|q|<1$, $|q|<|b|$, $|a|<1$. Analytic continuation implies it is valid for $|q|<1$, $|b/a|<|x|<1$. □

Acknowledgments. The author expresses his thanks to Professors Kazuhiko Aomoto and Yoshifumi Kato for useful suggestions.

REFERENCES

[1] G. E. ANDREWS, q-*Series: Their Development and Application in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra*, CBMS Regional Conference Series in Mathematics, 66, American Mathematical Society, Providence, RI, 1986.

[2] G. E. ANDREWS AND R. ASKEY, *A simple proof of Ramanujan's* ${}_1\Psi_1$ *summation*, Aequationes Math., 18 (1978), pp. 333–337.

[3] R. ASKEY, *Ramanujan's extensions of the gamma and beta functions*, Amer. Math. Monthly, 87 (1980), pp. 346–359.

[4] M. E. H. ISMAIL, *A simple proof of Ramanujan's* ${}_1\Psi_1$ *sum*, Proc. Amer. Math. Soc., 63 (1977), pp. 185–186.